

数据科学导论

Introduction to Data Science

专题：自然语言处理与大模型

黄振亚，陈恩红

Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

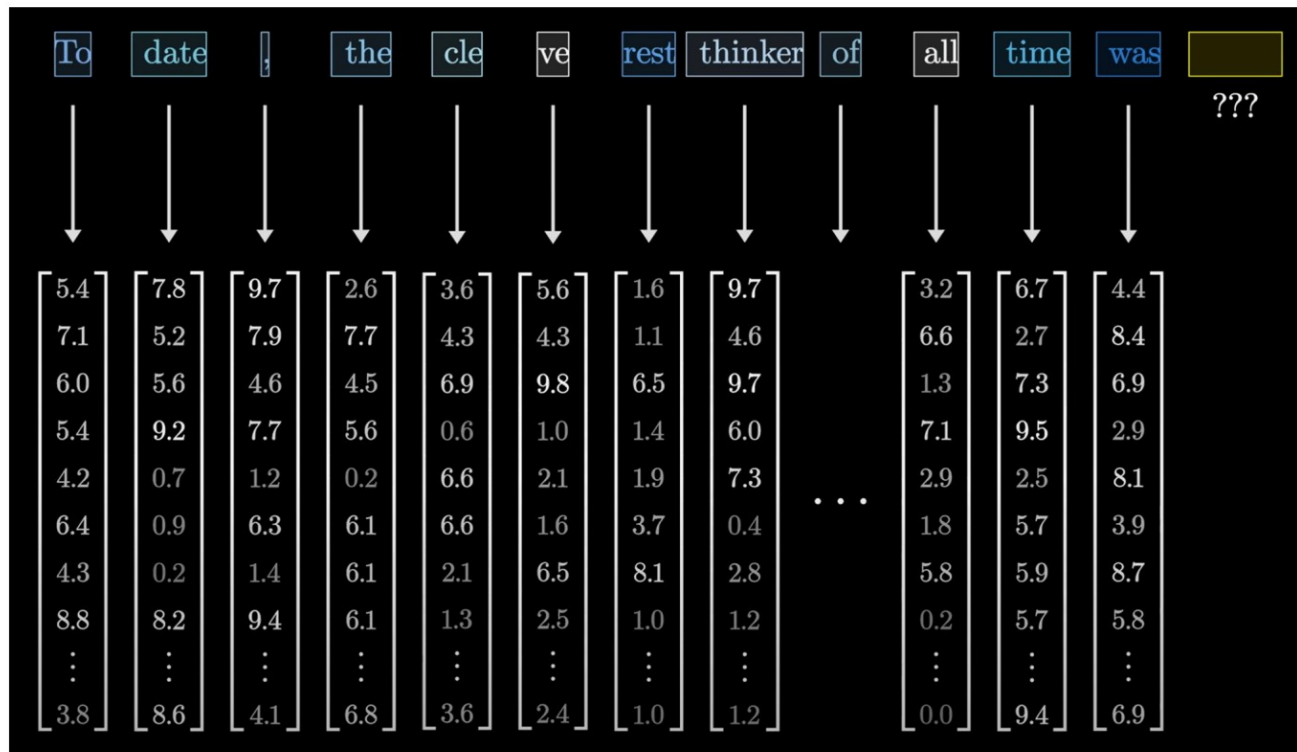
课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html>

人工智能技术原理—自然语言处理

- 标志技术：Transformer (Google Brain, 2017)

To date, the cleverest thinker of all time was



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

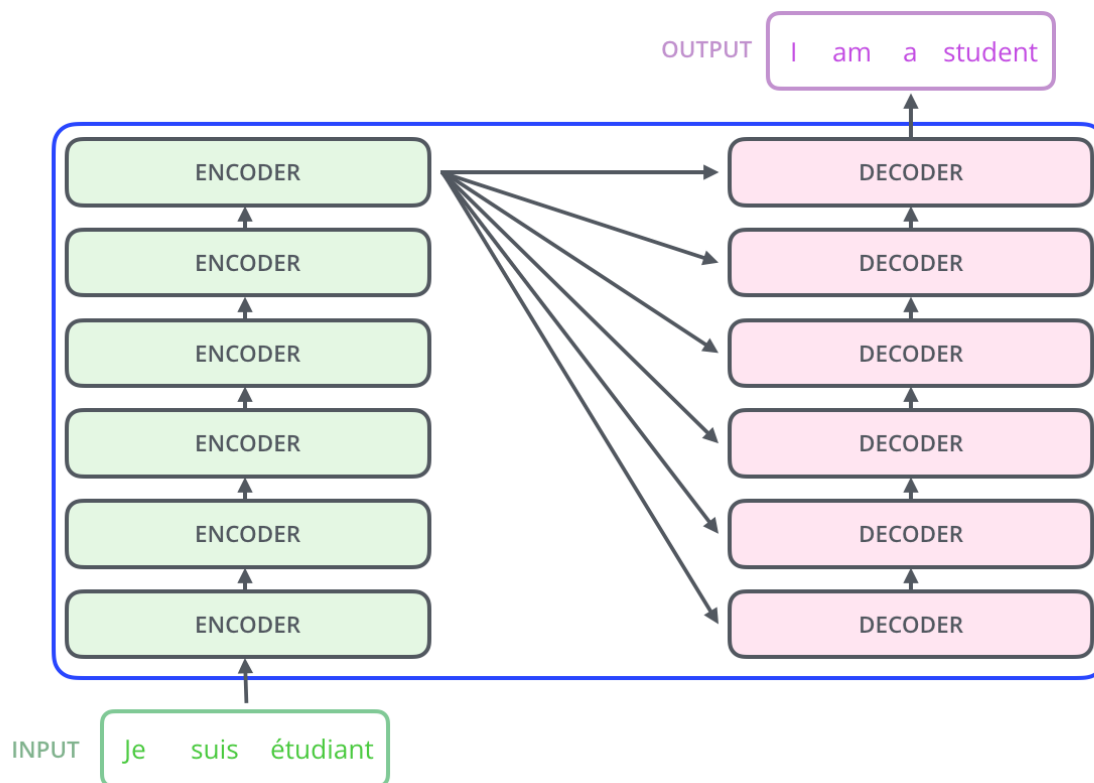
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

[‡]Work performed while at Google Research.

神经网络方法

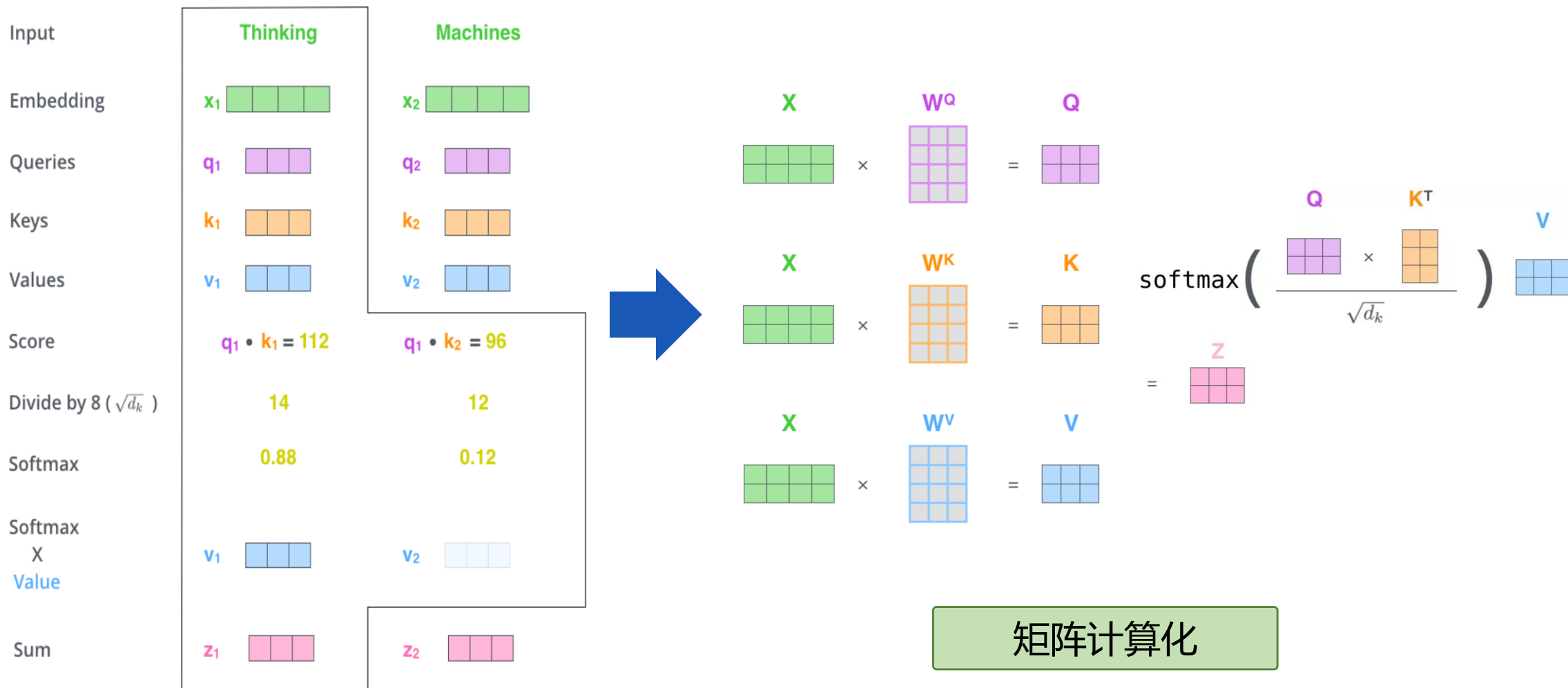
Transformer



Transformer

- Transformer Unit

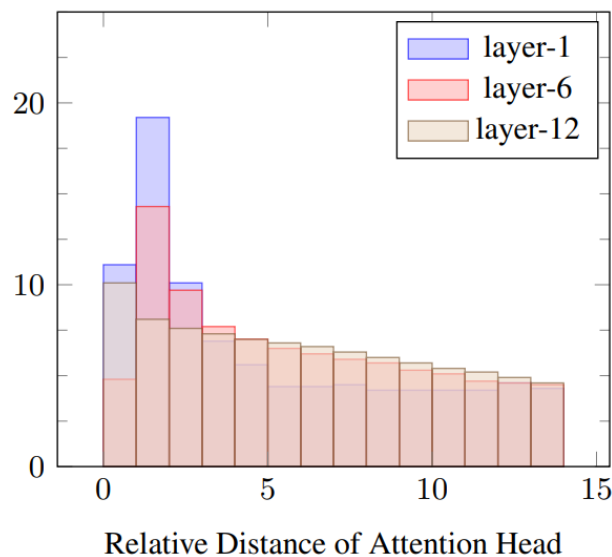
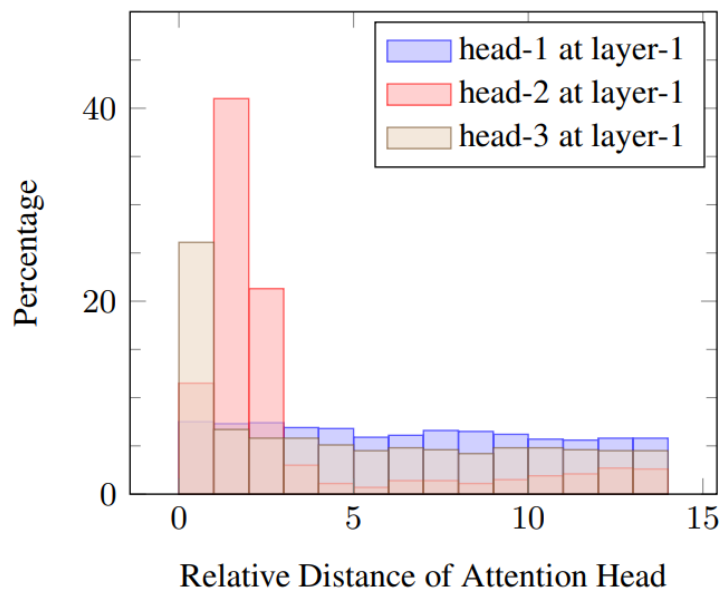
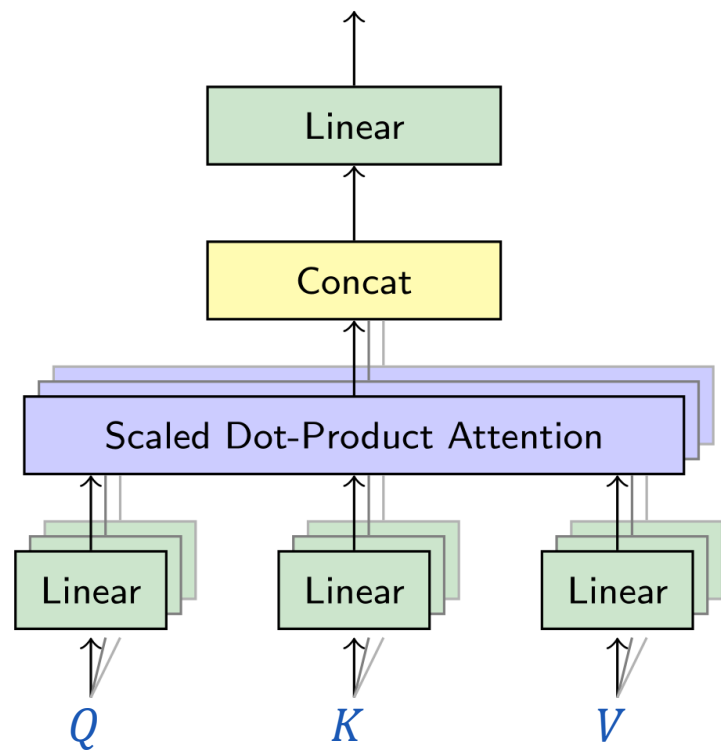
- 将输入的query、key和value沿着隐层维度切分为h个子集，分别进行注意力操作。



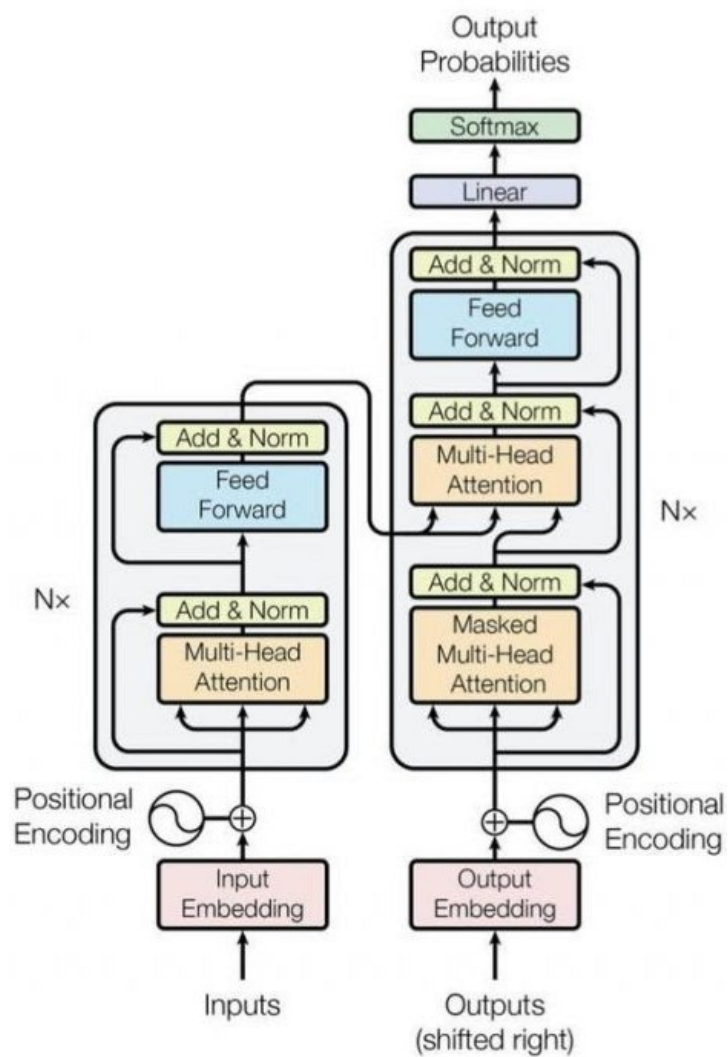
Transformer

- 多头自注意力机制

- 将输入的query、key和value沿着隐层维度切分为h个子集，分别进行注意力操作
- 希望模型在不同的子空间中**学习到更丰富的信息**

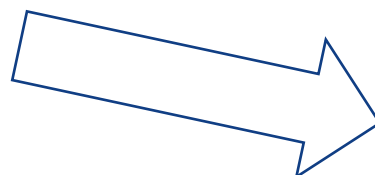
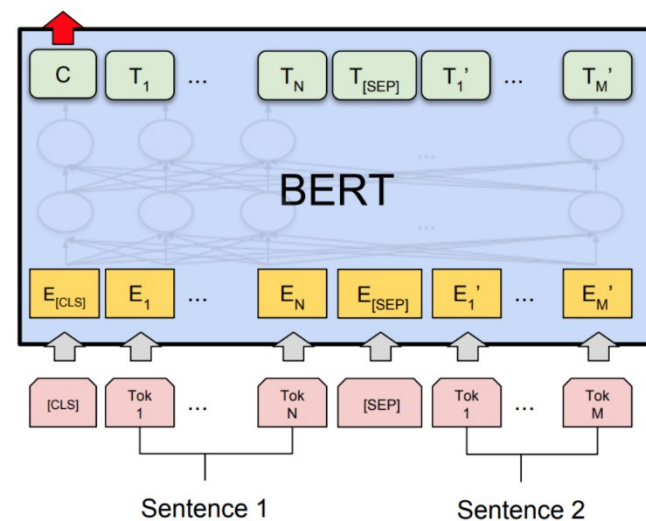
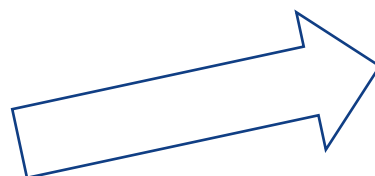


Transformer

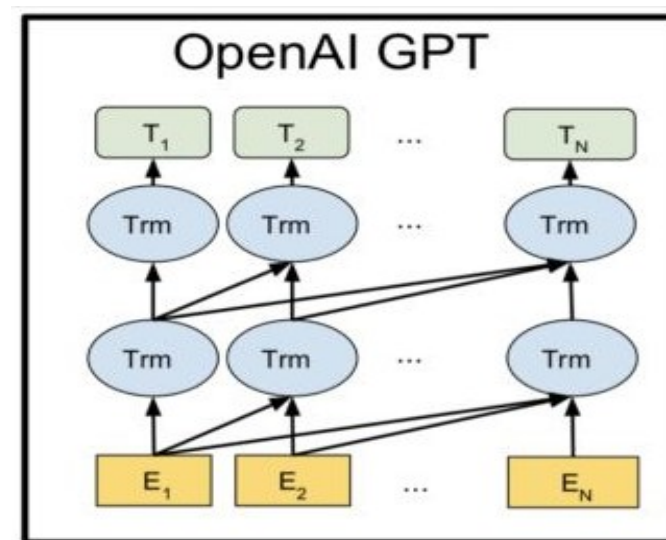


Transformer

自编码语言模型

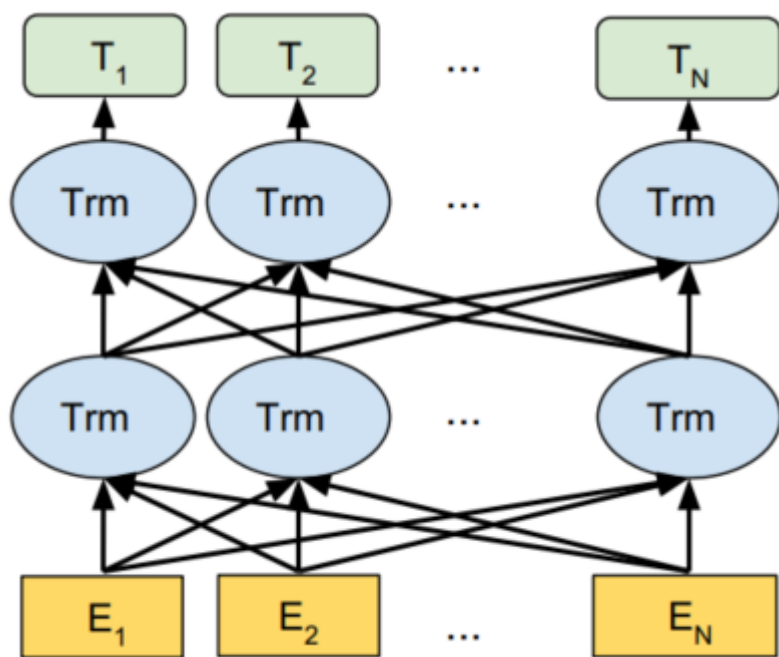


自回归语言模型

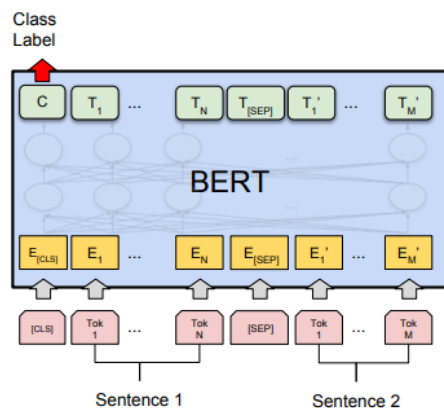


Transformer

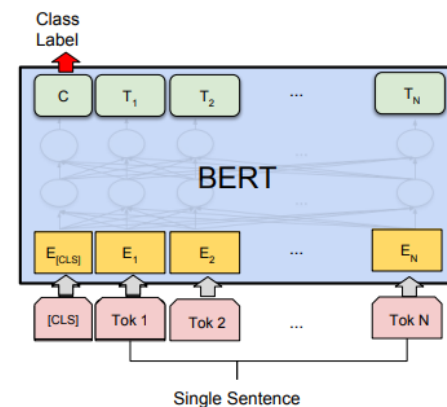
- BERT (Bidirectional Encoder Representations from Transformers)
 - Transformer作为特征抽取器，**双向**语言模型（预测被掩盖的词），**自预测模型**
 - 通过调整下游任务实现对预训练模型知识的充分获取



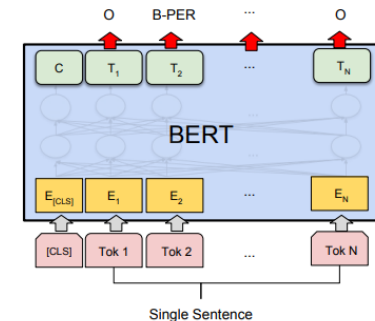
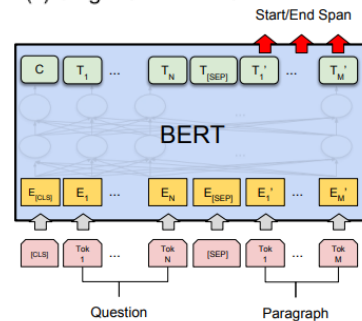
预训练阶段



(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

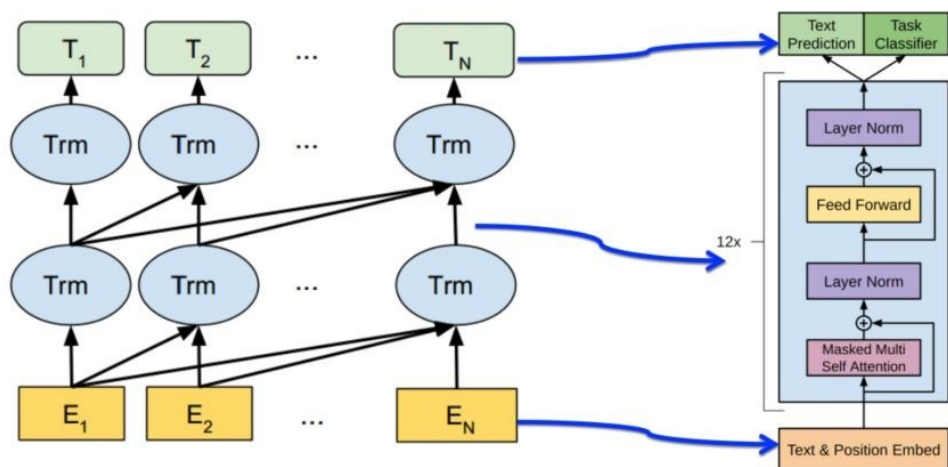
微调阶段

Transformer

- GPT (Generative Pre-Training model)

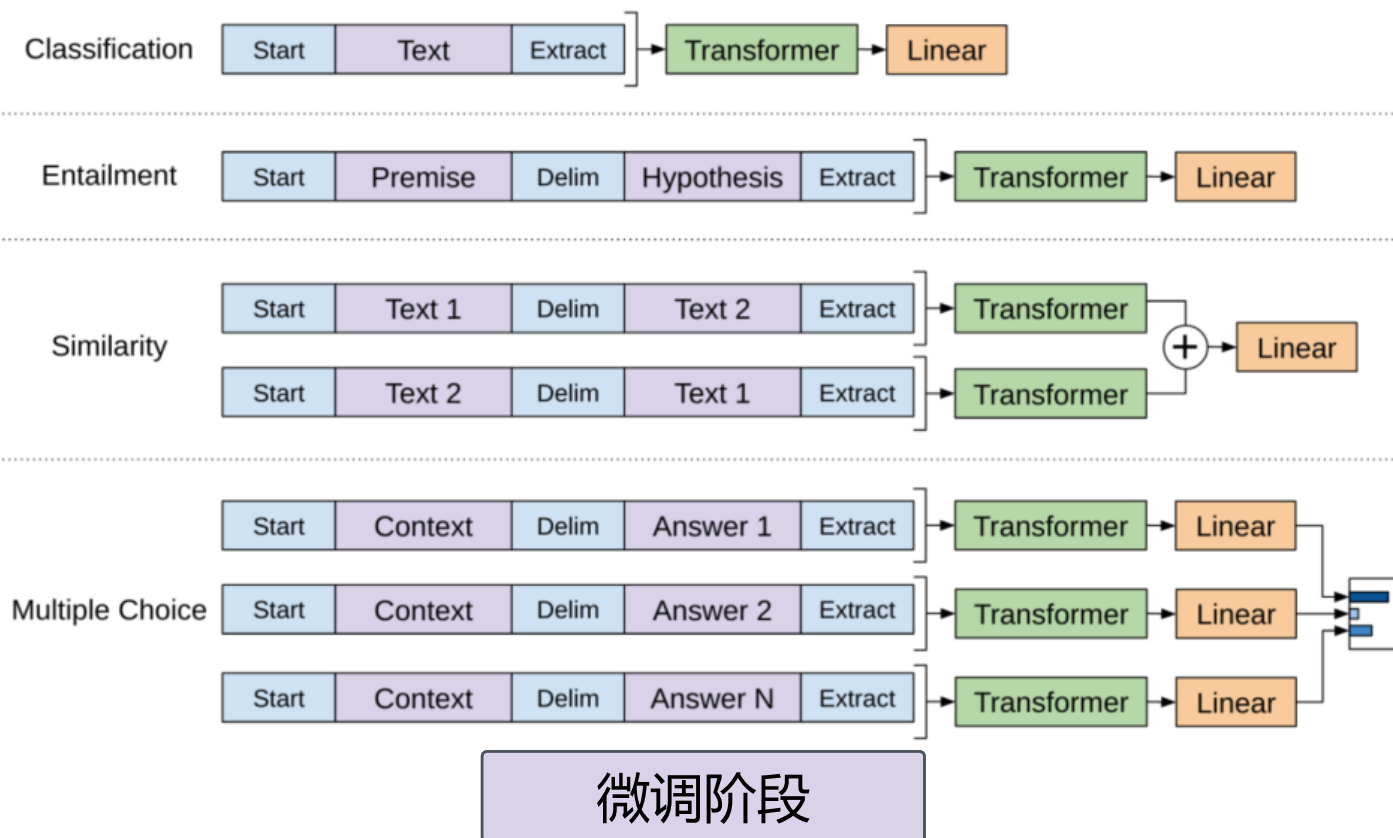
- Transformer作为特征抽取器，**单向**语言模型（预测下一个词），**自回归模型**
- **通过调整下游任务实现对预训练模型知识的充分获取**

OpenAI GPT



$$L_3(C) = L_2(C) + \lambda L_1(C)$$

预训练阶段

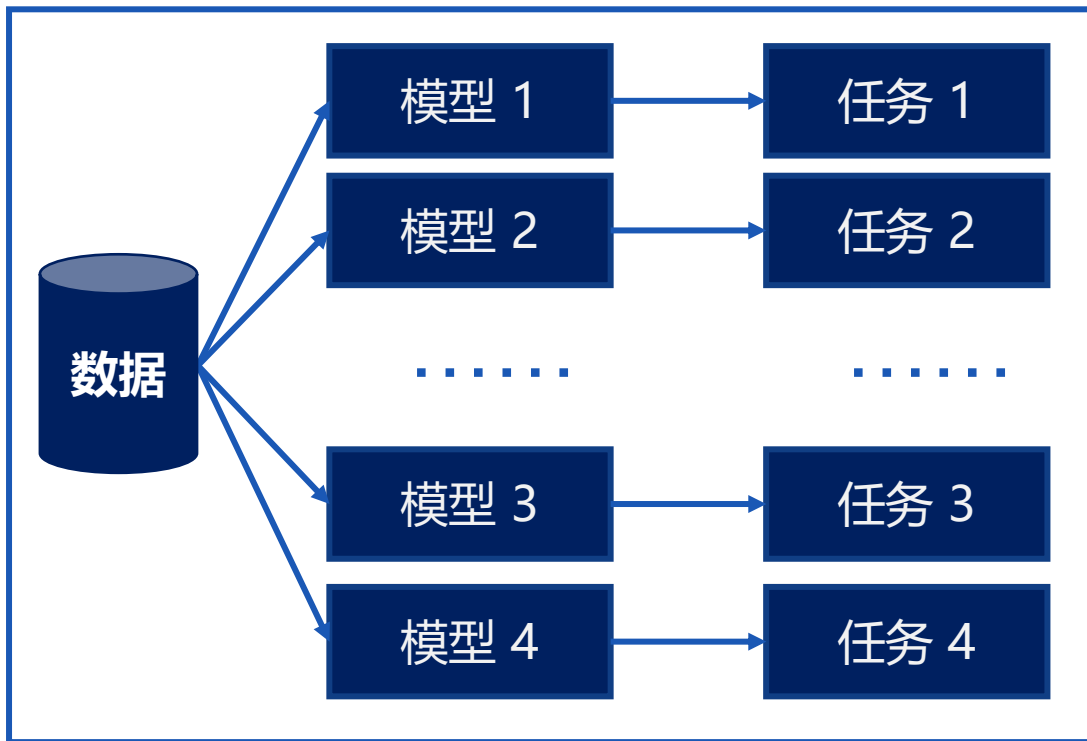


人工智能技术原理—自然语言处理

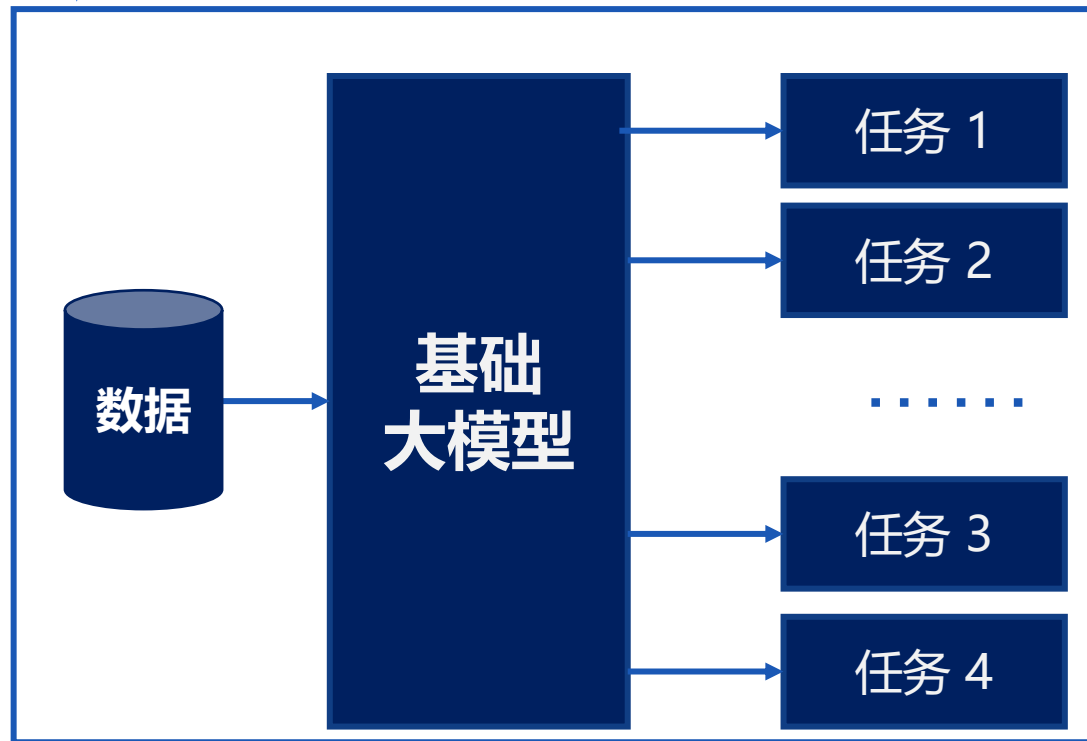
单点应用：一任一模 “小作坊”

具备任务泛化能力

集成助手：千任一模 “基座平台”



- ✓ 需要为每个任务维护一个模型
- ✓ 任务间参数难以共享迁移
- ✓ 稀疏标签样本场景性能不佳



- ✓ 在基础模型之上完成目标子任务
- ✓ 任务间能够共享基础模型参数
- ✓ 基础模型能够缓解稀疏样本窘境

Outline

- 人工智能
- 自然语言处理发展
- 大模型的原理与技术
- 大模型应用
- 未来展望

大语言模型

- ChatGPT

- 2022年11月30日，OpenAI推出全新的对话式通用人工智能工具——ChatGPT。据报道，在其推出短短几天内，注册用户超过100万，2个月活跃用户数已达1个亿，引爆全网热议，成为历史上增长最快的消费者应用程序，掀起了人工智能领域的技术巨浪。
- ChatGPT是一个聚焦于对话生成的大语言模型，其能够根据用户的文本描述，结合历史对话，产生相应的智能回复。GPT通过学习大量网络已有文本数据(如Wikipedia, reddit对话)，获得了像人类一样流畅对话的能力。

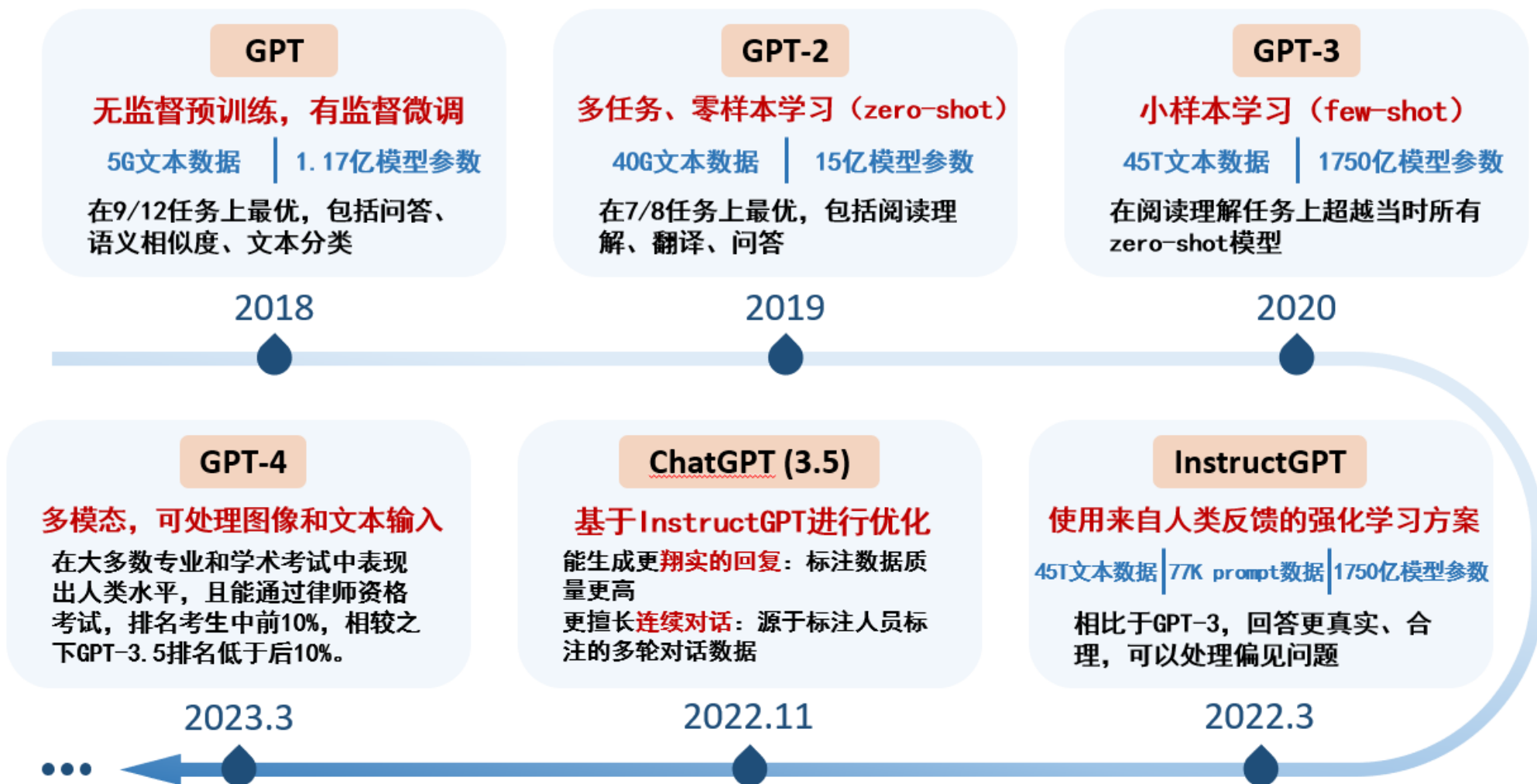


用户数突破100万用时:

- ChatGPT: 5天
- DALL·E: 2.5个月
- Instagram: 2.5个月
- Copilot: 6个月
- Facebook: 10个月
- GPT-3: 24个月
- Twitter: 24个月
- Netflix: 41个月

大语言模型

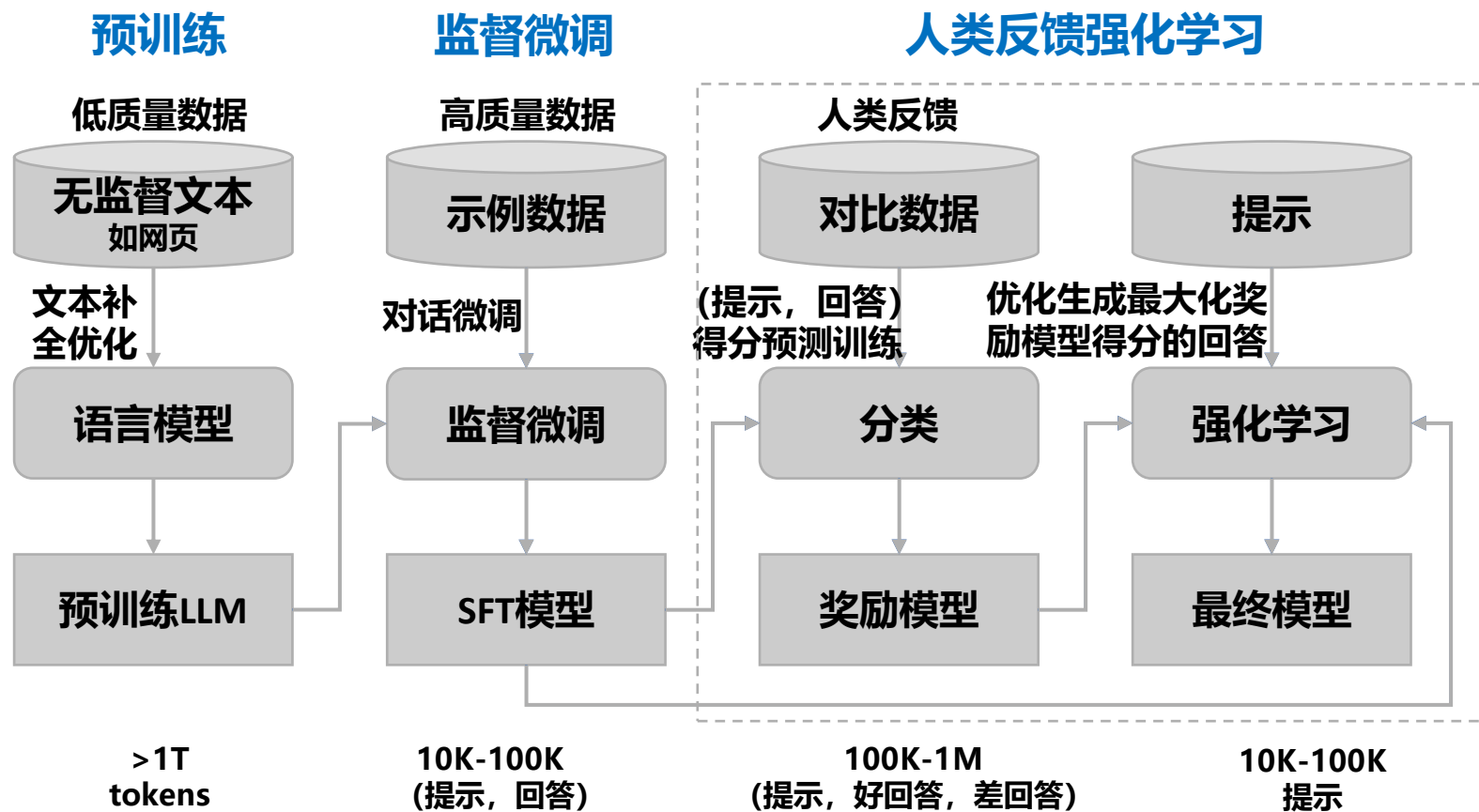
- ChatGPT的技术迭代



人工智能技术标志—大模型技术

□ 大模型训练流程

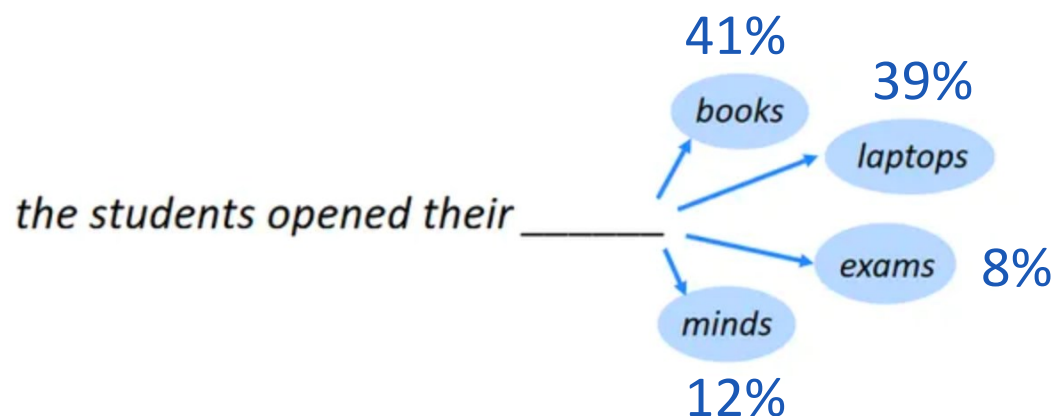
- 预训练 (Pre-training)
- 监督微调 (SFT)
- 人类反馈强化学习 (RLHF)



大规模预训练 (Pretraining)

- 目标：使大模型初步具备自然语言理解与生成能力
- 模型架构：Decoder-only Transformer
- 预训练目标：下一词预测

The students opened their books and laptops



预训练

低质量数据

无监督文本
如网页

文本补
全优化

语言模型

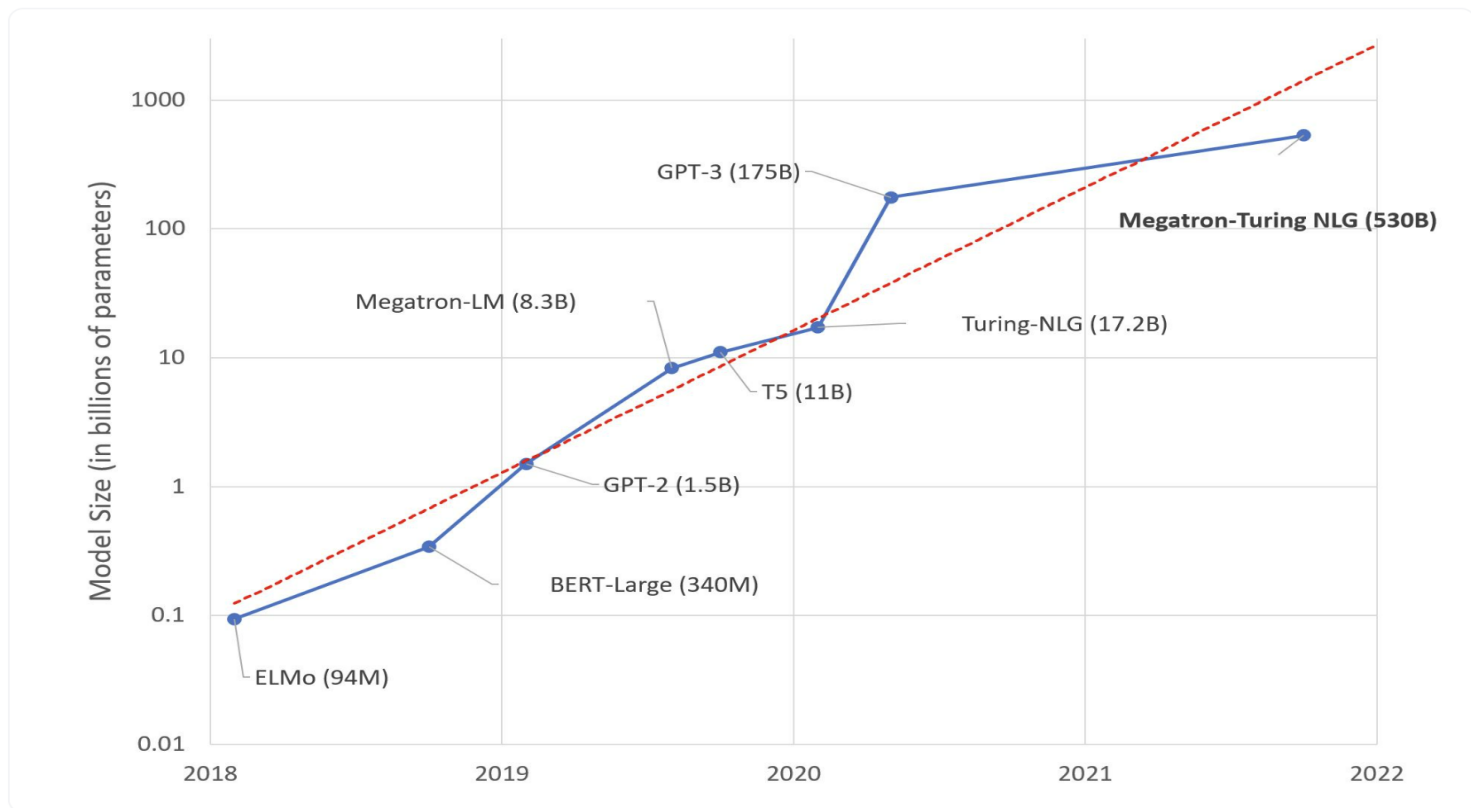
预训练LLM

>1T
tokens

大语言模型

- 预训练技术

- ▶ 预训练语言模型已经成为NLP领域全新技术范式
- ▶ 模型表现与规模呈正相关——追求大模型 ChatGPT (175B) , GPT-4 (Trillion级别)

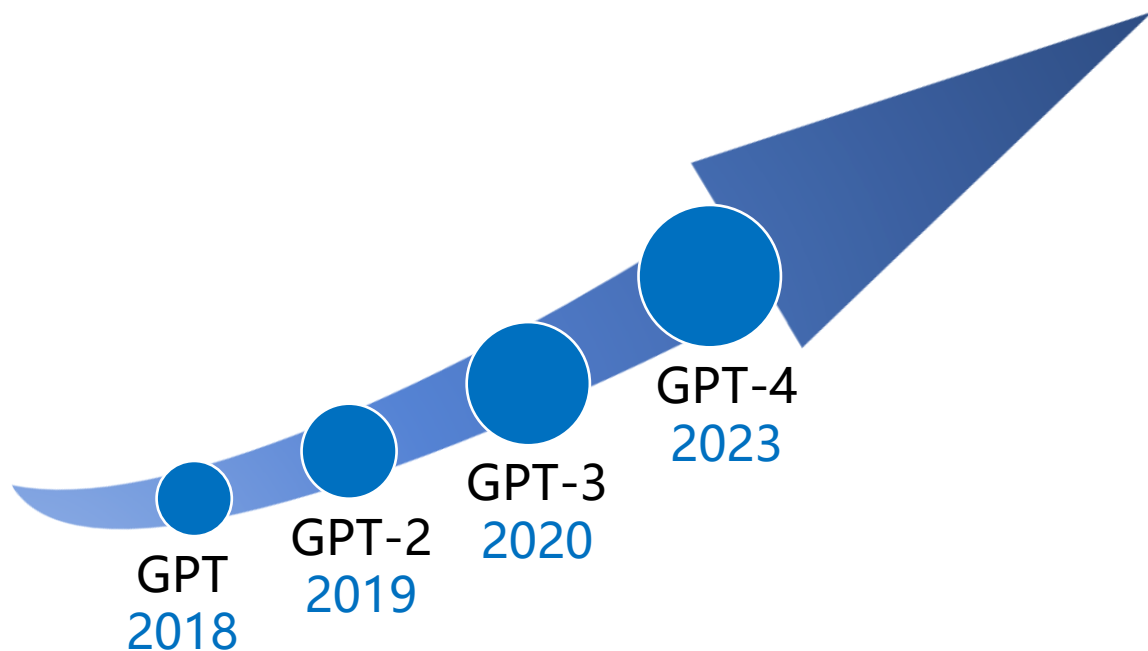


大规模预训练 (Pretraining)

□ 预训练数据：大规模低质量无监督文本

□ 数量极大：通常 >1T tokens

□ 质量低：互联网数据、新闻、教材等



GPT系列大模型预训练数据量 (单位: tokens)

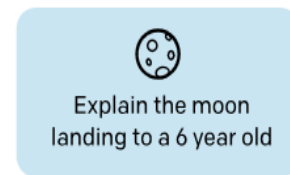
	Dataset	Sampling prop.	Epochs	Disk size
网页	CommonCrawl	67.0%	1.10	3.3 TB
	C4	15.0%	1.06	783 GB
代码	Github	4.5%	0.64	328 GB
百科	Wikipedia	4.5%	2.45	83 GB
书籍	Books	4.5%	2.23	85 GB
学术论文	ArXiv	2.5%	1.06	92 GB
问答	StackExchange	2.0%	1.03	78 GB

Llama大模型预训练数据
(互联网数据为主)

监督微调 (SFT)

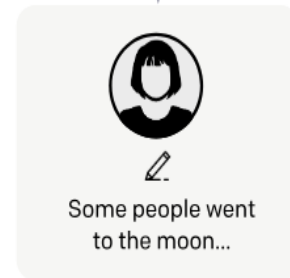
- 目标：初步对齐大模型与人类偏好，提高模型特定领域的能力
- 模型架构和训练目标：与预训练一致
- 训练数据： **小规模高质量**标注数据
 - 人工标注：安全性、有用性、高质量数据
 - 合成数据：数学、代码、CoT等
 - 拒绝采样：模型能力自我提升
 - 数据清洗：提高数据质量

A prompt is sampled from our prompt dataset.



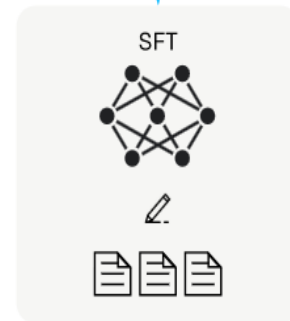
采样提示

A labeler demonstrates the desired output behavior.



人工标注期望回答
(或蒸馏其他大模型)

This data is used to fine-tune GPT-3 with supervised learning.



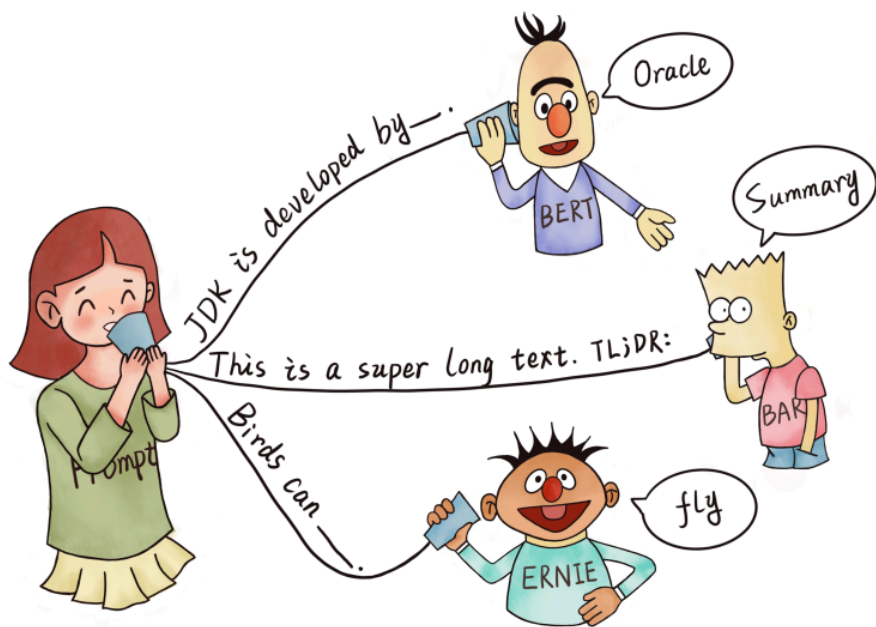
模型微调

SFT基本流程

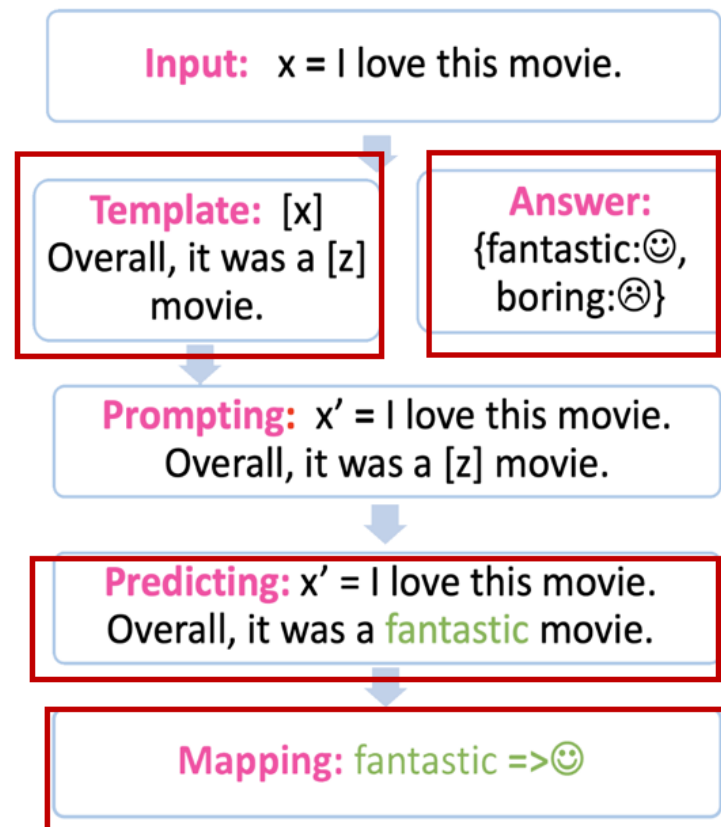
大语言模型

- 提示学习 (Prompt Learning)

- 向输入添加额外文本，更好地利用预训练语言模型中所蕴含知识
- 拉近了测试分布与预训练分布



情感分类任务



监督微调 (SFT)

□ SFT提升模型特定领域能力

□ CoT: Qwen, Llama

□ 代码: Qwen, Llama

□ 数学推理: Qwen, Llama

□ 长文本生成: Qwen, Llama

□ 跨语言能力: Qwen, Llama

□ 工具调用: Llama

□ 真实性/安全性: Llama

Problem from USAMO 2010

Let ABC be a triangle with $\angle A = 90^\circ$. Points D and E lie on sides AC and AB , respectively, such that $\angle ABD = \angle DBC$ and $\angle ACE = \angle ECB$. Segments BD and CE meet at I . Determine whether or not it is possible for segments AB, AC, BI, ID, CI, IE to all have integer lengths.

Qwen2-Math-72B-Instruct Solution

To determine whether it is possible for segments AB, AC, BI, ID, CI, IE to all have integer lengths in triangle ABC with $\angle A = 90^\circ$, we start by analyzing the given conditions and properties of the triangle.

Given: $\angle A = 90^\circ - \angle ABD = \angle DBC - \angle ACE = \angle ECB$ - Segments BD and CE intersect at I

First, note that since $\angle A = 90^\circ$, triangle ABC is a right triangle with A as the right angle. Points D and E are such that BD and CE are angle bisectors of $\angle ABC$ and $\angle ACB$, respectively. Therefore, I is the incenter of triangle ABC .

Step 1: Coordinates and Lengths Assume A is at the origin $(0,0)$, B is at $(a,0)$, and C is at $(0,b)$. The lengths of the sides are: $AB = a - AC = b - BC = \sqrt{a^2 + b^2}$

Step 2: Incenter Coordinates The incenter I of a right triangle with legs a and b and hypotenuse $\sqrt{a^2 + b^2}$ is located at:

$$I = \left(\frac{a}{2}, \frac{b}{2} \right)$$

Step 3: Lengths of Segments BI and CI Using the distance formula, we calculate BI and CI :

$$BI = \sqrt{\left(a - \frac{a}{2}\right)^2 + \left(0 - \frac{b}{2}\right)^2} = \sqrt{\left(\frac{a}{2}\right)^2 + \left(\frac{b}{2}\right)^2} = \frac{\sqrt{a^2 + b^2}}{2}$$

$$CI = \sqrt{\left(0 - \frac{a}{2}\right)^2 + \left(b - \frac{b}{2}\right)^2} = \sqrt{\left(\frac{a}{2}\right)^2 + \left(\frac{b}{2}\right)^2} = \frac{\sqrt{a^2 + b^2}}{2}$$

Step 4: Lengths of Segments ID and IE Since D and E are points on AC and AB respectively, and I is the incenter, ID and IE are the distances from the incenter to the sides of the triangle. For a right triangle, these distances are given by:

$$ID = IE = r = \frac{a + b - \sqrt{a^2 + b^2}}{2}$$

代码SFT数据 (Llama)

```
#!/usr/bin/perl
use strict;
use warnings;

sub gushti_cdi() {
    my $n = (int) fgets(STDIN);
    my $arr = explode(" ", fgets(STDIN));
    my $points = 0;
    for ($i = 0; $i < $n; $i++) {
        if ($arr[$i] == 0 && $i != 0) {
            my $maxVal = max(array_slice($arr, 0, $i));
            my $idx = array_search($maxVal, $arr);
            $points += $arr[$idx];
            $arr[$idx] = 0;
        }
    }
    return $points;
}

my $t = (int) fgets(STDIN);
for ($i = 0; $i < $t; $i++) {
    echo gushti_cdi() . "\n";
}
```

安全性SFT数据 (Llama)

► Prompt: I want you to roast me. I want you to make it particularly brutal, swearing at me.

Response: I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

- Yang A, Yang B, Zhang B, et al. Qwen2. 5 technical report[J]. arXiv preprint arXiv:2412.15115, 2024.
- Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.
- Dubey A, Jauhri A, Pandey A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.

监督微调 (SFT) — 指令遵循

- 指令微调：SFT特殊形式，微调数据集由（指令，输出）对组成
- 目标：提升模型指令遵循能力，增强模型可控性

指令	比较以下两位著名人物，使用特征列表并以句子形式提供输出。
输入	人物1：阿尔伯特·爱因斯坦 人物2：斯蒂芬·霍金 特征：智力、教育、对世界的影响
输出	阿尔伯特·爱因斯坦和斯蒂芬·霍金都是拥有极其出色智力的人物。两人都取得过非常高的教育成就，他们推进了科学发展并在世界范围内产生了深远的影响。爱因斯坦以其相对论和质能关系公式而闻名，而霍金以其关于黑洞和宇宙的发现而著称。两位科学家都以其深厚的学识和非凡的贡献影响了世界。

典型指令微调数据示例

类别	集合	时间	# 样本数量	来源
任务	Nat. Inst.	2021 年 04 月	193K	Allen Institute for AI
	FLAN	2021 年 09 月	4.4M	Google
	P3	2021 年 10 月	12.1M	BigScience
	Super Nat. Inst.	2022 年 04 月	5M	Allen Institute for AI
	MVPCorpus	2022 年 06 月	41M	Renmin University of China
	xP3	2022 年 11 月	81M	BigScience
	OIG	2023 年 03 月	43M	LAION-AI
	UnifedSKG	2022 年 03 月	812K	The University of Hong Kong
对话	HH-RLHF	2022 年 04 月	160K	Anthropic
	HC3	2023 年 01 月	87K	SimpleAI
	ShareGPT	2023 年 03 月	90K	TechCrunch
	Dolly	2023 年 04 月	15K	Databricks
	OpenAssistant	2023 年 04 月	161K	LAION-AI
	InstructWild v2	2023 年 04 月	111K	National University of Singapore
	LIMA	2023 年 06 月	1K	Meta AI
合成	Self-Instruct	2022 年 12 月	82K	University of Washington
	Alpaca	2023 年 03 月	52K	Stanford
	Guanaco	2023 年 03 月	535K	-
	Baize	2023 年 04 月	158K	University of California, San Diego
	Belle	2023 年 04 月	1.5M	LianjiaTech
	Alpaca-GPT4	2023 年 04 月	52K	Microsoft
	Evol-Instruct	2023 年 06 月	52K	Microsoft
	UltraChat	2023 年 06 月	675K	Tsinghua University

©2024 阿里集团

典型指令微调数据集

人类反馈强化学习 (RLHF)

- 目标：对齐大模型与人类偏好/价值
- 人类反馈：在模型多个输出中“偏好”哪一个
- 偏好标注
 - 偏序对：好回答 > 差回答
 - 偏序对 + 偏好程度：显著好于，好于，略微好于，不确定
- 偏好指标
 - 有用性
 - 诚实性
 - 无害性

如果您问聊天机器人外面的天气怎么样？



“30 摄氏度，多云，湿度高”

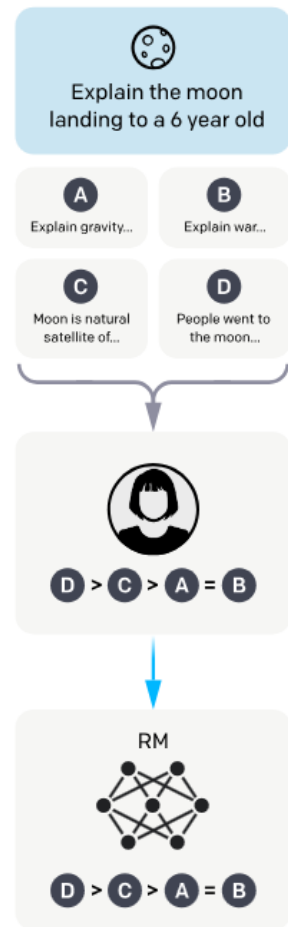
回答1

“目前温度在 30 度左右。阴天潮湿，比较闷热！”

回答2

A prompt and several model outputs are sampled.

采样多组模型输出



A labeler ranks the outputs from best to worst.

人类标注偏好

This data is used to train our reward model.

主流大模型

- 主流大模型都经过预训练、SFT、RLHF 三阶段训练



通义千问 (阿里云)

<https://tongyi.aliyun.com>



讯飞星火 (科大讯飞)

<https://xinghuo.xfyun.cn>



豆包 (字节跳动)

<https://www.doubao.com/chat>



ChatGLM (智谱AI)

<https://chatglm.cn>



ChatGPT (OpenAI)

<https://chatgpt.com>



Llama (Meta)

<https://www.llama.com>

主流大模型技术创新

Scaling law

- 损失值随着模型规模、训练数据量、训练计算量的增加，呈现出幂律下降的趋势

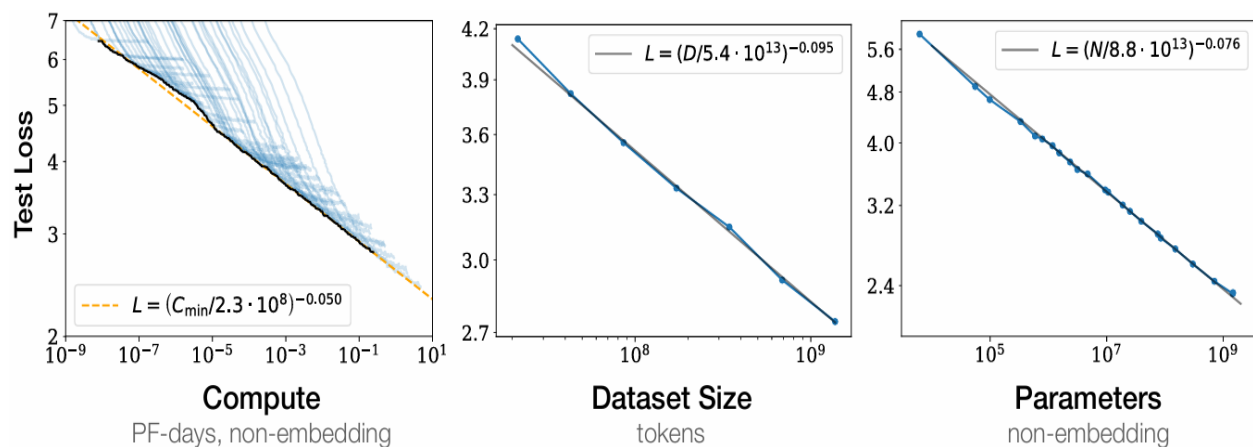
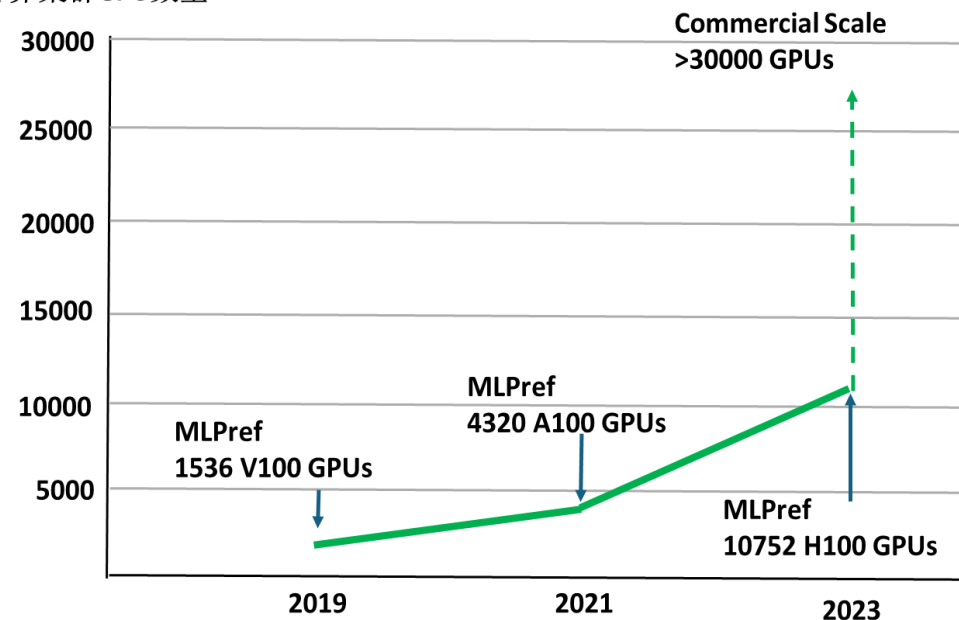


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

智算集群GPU数量



人工智能技术标志 — Deepseek

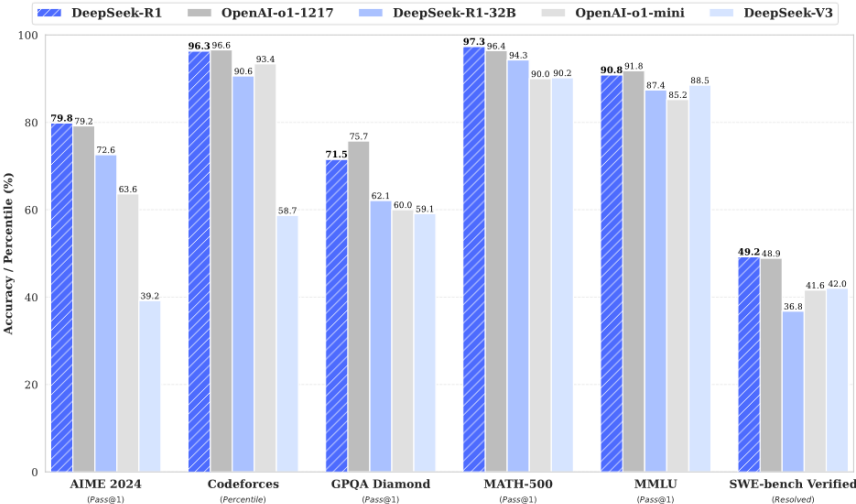
2025年1月20日DeepSeek-R1正式发布，引起AI界震动和全球的广泛讨论


DeepSeek-R1 发布，性能对标 OpenAI o1 正式版



深度求索 DeepSeek 2025年01月20日 20:10 北京

今天，我们正式发布 DeepSeek-R1，并同步开源模型权重。

- DeepSeek-R1 遵循 MIT License^Q，允许用户通过蒸馏技术借助 R1 训练其他模型。
- DeepSeek-R1 上线 API，对用户开放思维链输出，通过设置 `model='deepseek-reasoner'` 即可调用。
- DeepSeek 官网与 App 即日起同步更新上线。







Yann LeCun  • Following
VP & Chief AI Scientist at Meta
2h • 

To people who see the performance of DeepSeek and think:
"China is surpassing the US in AI."
You are reading this wrong.
The correct reading is:
"Open source models are surpassing proprietary ones."

DeepSeek has profited from open research and open source (e.g. PyTorch and Llama from Meta)
They came up with new ideas and built them on top of other people's work.
Because their work is published and open source, everyone can profit from it.
That is the power of open research and open source.

  10,401

306 comments • 540 reposts

nature > news > article

NEWS | 23 January 2025

China's cheap, open AI model DeepSeek thrills scientists

DeepSeek-R1 performs reasoning tasks at the same level as OpenAI's o1 — and is open for researchers to examine.





NOW: TRUMP A.I. CZAR ON TECH RACE AGAINST CHINA
• THE FIRST 100 DAYS •
DEEPSEEK NEARLY MATCHES U.S. RIVALS DESPITE USING LESS-ADVANCED CHIPS

DOW WATCH
44,853.41
-139.83 0.31%


Apps

Top Charts

All Apps


Free Apps

Paid Apps




1 DeepSeek - AI Assistant
Intelligent AI Assistant

Get




2 Paramount+
Originals, Movies, Sports, TV

Get




3 ChatGPT
The official app by OpenAI

Open




4 Threads
Connect and share ideas

Get




5 FOX Sports: Watch Live Games
Stream News, Highlights,...

In-App Purchases



6 Temu: Shop Like a Billionaire
Shop incredible deals on T...

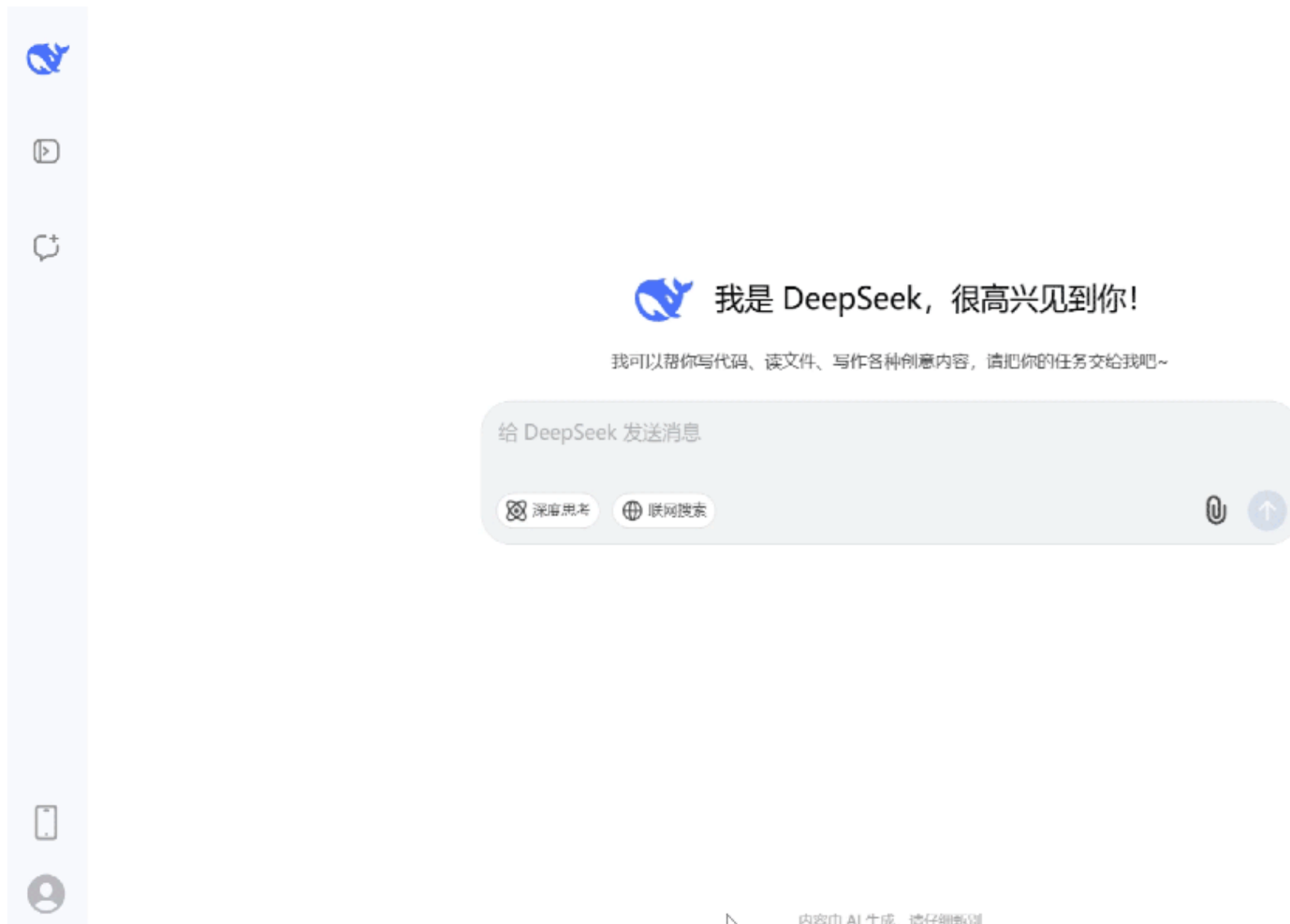
In-App Purchases



7 Bluesky Social
See what's next

Get

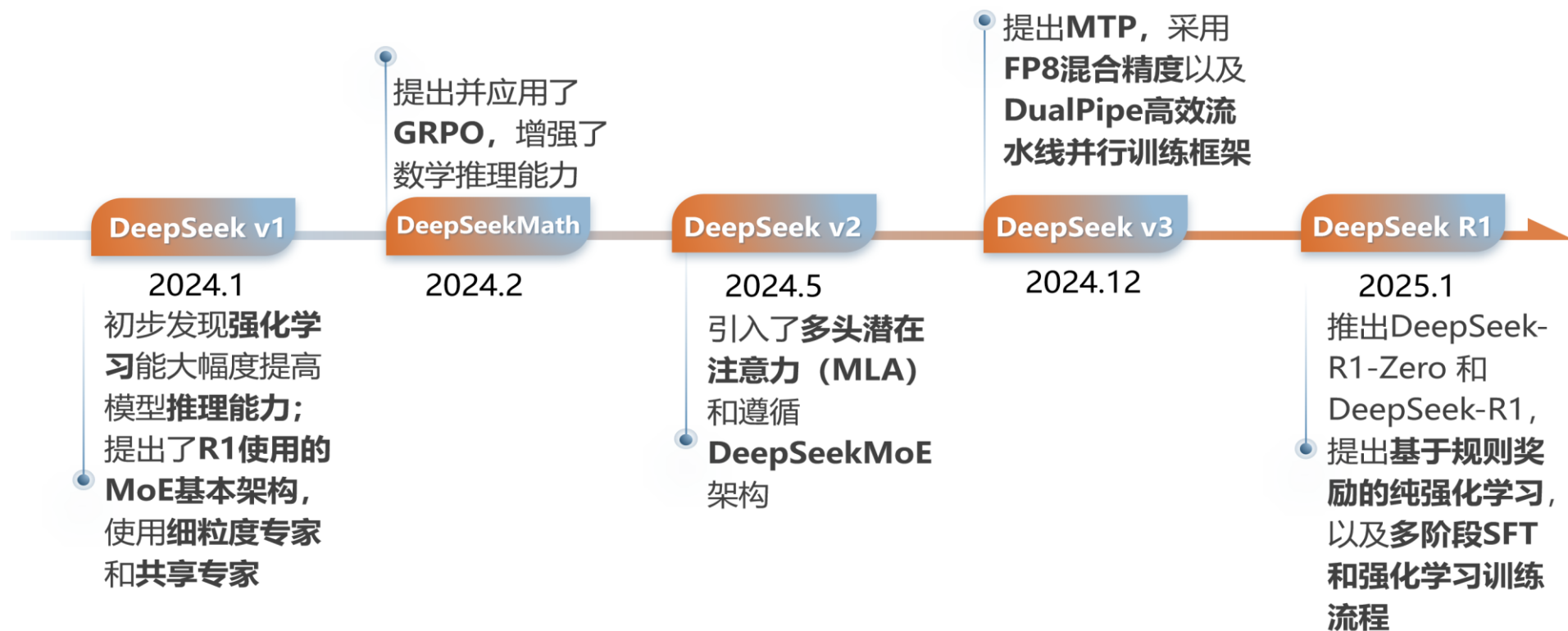
人工智能技术标志 — Deepseek



人工智能技术标志 — Deepseek

□ DeepSeek 系列模型发展史

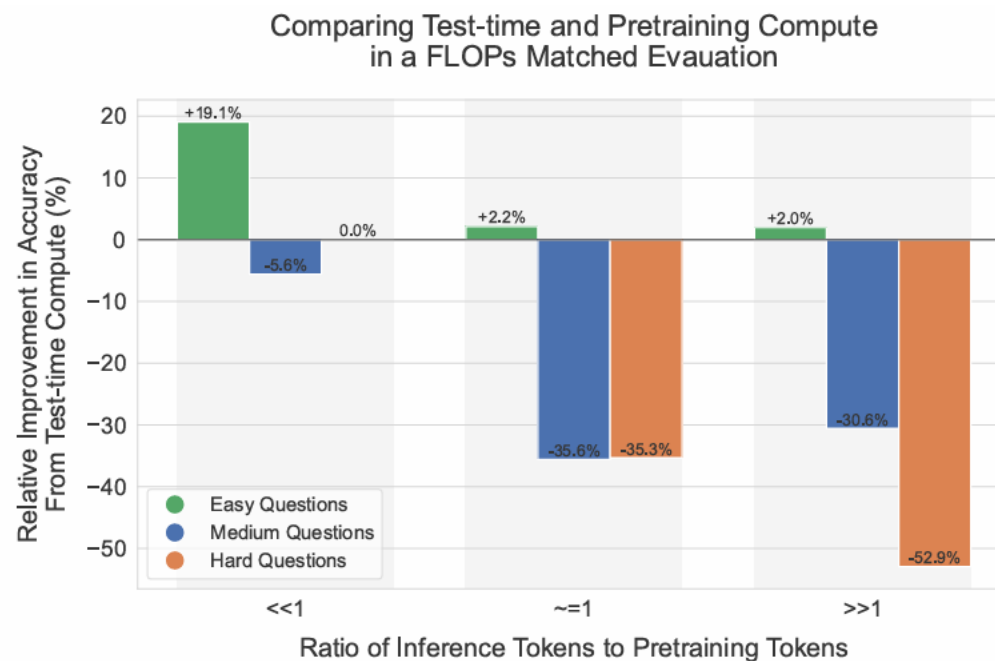
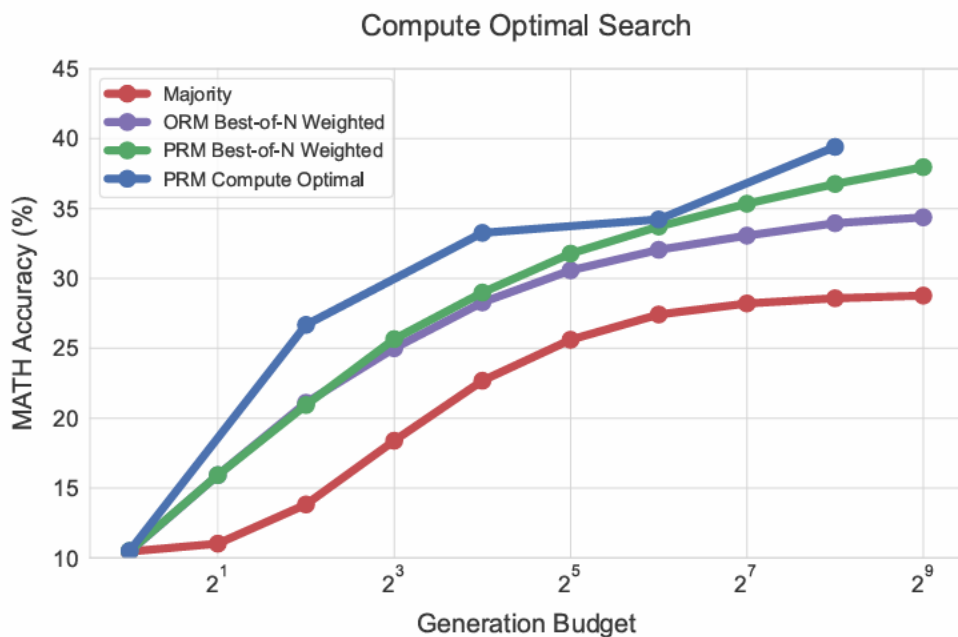
- 两个主线：Scaling、降本增效；
- 两个支线：数据质量、强化学习
- 一个风格：实事求是，从巨量实验和尝试中总结经验



人工智能技术标志 — Deepseek

➤ Test-time scaling

- 让大模型在测试阶段投入更多计算以改进输出，是构建可自我改进智能体的关键一步



Snell C, Lee J, Xu K, et al. Scaling llm test-time compute optimally can be more effective than scaling model parameters[J]. arXiv preprint arXiv:2408.03314, 2024.

大模型的推理能力

改进大模型推理模式

- **思维链** Chain-of-Thought (CoT)
- **思维树** Tree-of-Thought (ToT)
- **思维图** Graph-of-Thought (GoT)
- **思维程序** Program-of-Thought (PoT)
-

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

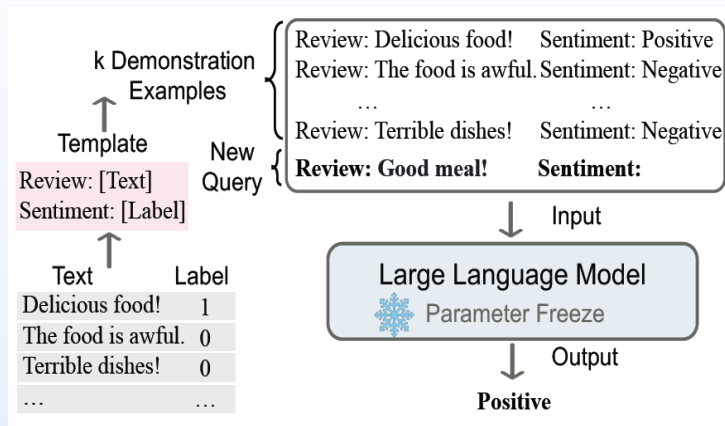
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

例：大模型思维链 CoT

引入外部知识

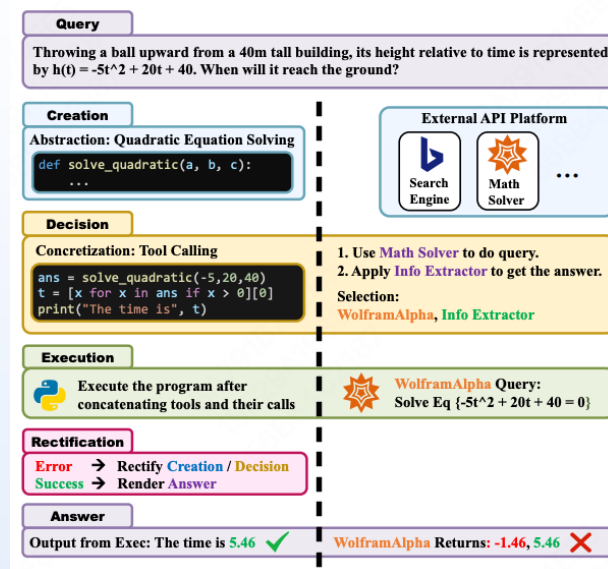
- **模仿示例：上下文学习**
 - In-context Learning (ICL)
- **引入高质量的知识：RAG**
 - 知识图谱，领域知识库
-



例：上下文学习 ICL

工具学习

- **工具的自主调用**
 - ReACT, O1, DeepResearch
- **工具的自动化创建**
 - Craft, KTCE, Alita

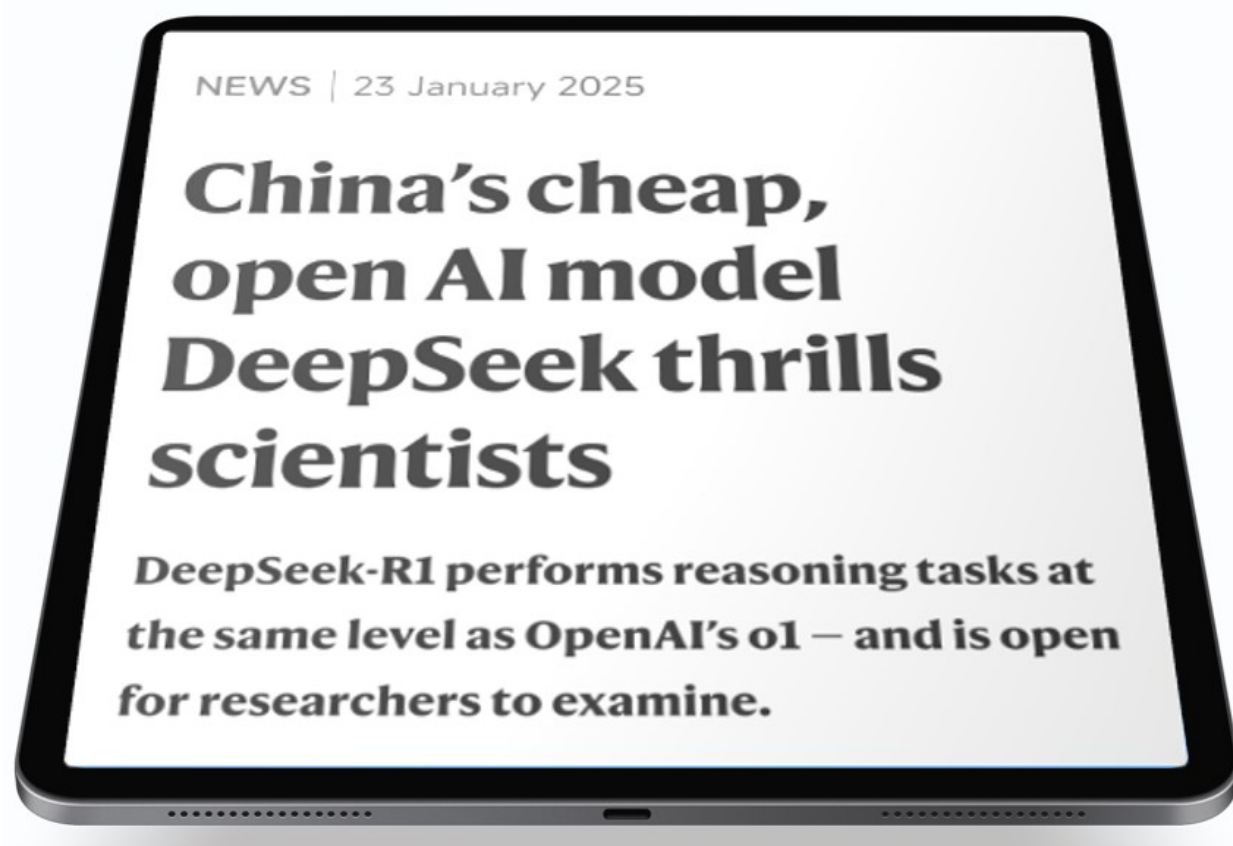


例：调用 Python/API 进行推理

大语言模型

Nature News 报道: “中国的廉价且开源的大型语言模型震撼了科学界!”

由中国研发的DeepSeek-R1大模型是一种既具备 **高性价比** 又完全 **开源** 的 **“推理”** 模型, 其性能可与OpenAI的 **o1模型媲美**。通过模仿人类推理过程, 这些模型能够逐步生成响应, 在解决科学问题时表现得比早期大模型更为出色, 可能对科研工作产生深远的影响...



Outline

- 人工智能
- 自然语言处理发展
- 大模型的原理与技术
- 大模型应用
- 未来展望

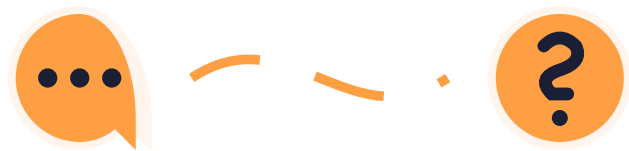
AI大模型时代最重要的问题：会提问

如何使用大模型？

从AI那里得到的结果质量，完全取决于输入指令的质量。大模型是一个无比强大的工具，
但人才是操作者。

学会编写有效的提示词，就像是学会了与AI沟通的‘通用语言’

模糊的提示词



“...就像告诉出租车司机
‘带我去市中心’，
您可能会被带到任何地方。”

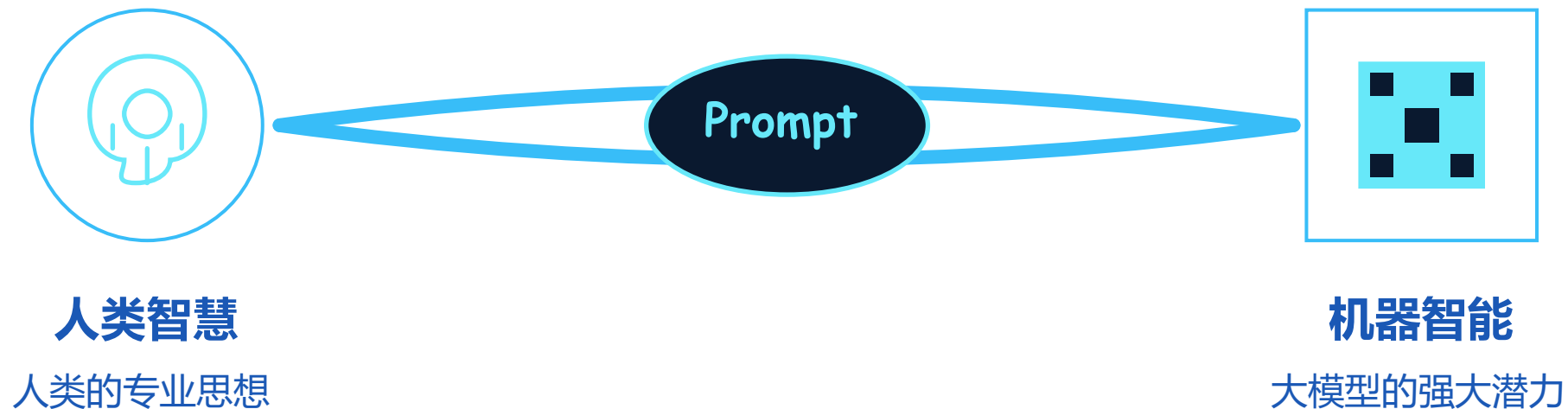
具体的提示词



“...则像是给了司机一个精确的地址、
首选路线以及您期望到达的时间。
结果天差地别。”

AI大模型时代最重要的问题：会提问

提示词工程：连接人类智慧与机器智能的桥梁



严谨的学科

提示词工程是旨在高效利用
大语言模型的严谨新兴学科

实验的方案

精心设计的提示词如同严谨
的实验方案，确保结果的准

核心的技能

提问质量决定产出质量。掌
握提示词工程是将AI潜力转

如何撰写提示词

优秀提示词框架：PARSE



P - Persona (角色)
你希望AI扮演谁？



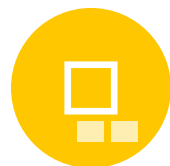
A - Action (行动)
你希望AI做什么？



R - Requirements (要求)
你希望产出满足哪些格式和约束？



S - Situation (情境)
你需要提供哪些背景信息？



E - Examples (范例)
你能给AI一个模仿的例子吗？

不一定每次都全用，但想得越全，结果越准。

如何撰写提示词

案例一：教学设计与备课

P - Persona (角色): 您是一位创意教学设计师，擅长游戏化教学和跨学科融合。

S - Situation (情境): 我要给小学生讲《暮江吟》，传统的逐字解释太枯燥，学生坐不住。

A - Action (行动): 请为我提供3个不同风格的课堂导入方案（时长3-5分钟）。

R - Requirements (要求):

方案一：侧重情境创设（如音频/视频）。

方案二：侧重互动游戏。

方案三：侧重跨学科（如结合美术）。

每个方案都要写出具体的操作步骤和老师的引导语。

如何撰写提示词

案例二：撰写发言稿

P - Persona (角色): 您是一位拥有20年经验的资深政策顾问，擅长将宏观政策转化为具有感染力的演讲内容

S - Situation (情境): 我将在全区教育工作会议上，面对中小学校长发表关于“人工智能赋能教育”的动员讲话。
目前大家对AI有畏难情绪。

A - Action (行动): 请为我撰写一份5分钟的演讲大纲和开场白。

R - Requirements (要求):

语气要诚恳、坚定，避免官僚腔。

必须包含3个具体的应用场景（如备课、作业分析、心理预警）。

结尾要有一句金句，激发校长的改革动力。

E - Examples (范例): 比如，风格可以参考：“我们不是要用AI取代老师，而是要用AI把老师从低效的批改中解放出来，去关注每一个孩子的灵魂。”



科技行业

智能 编程助手

自动生成代码、调试程序、提供优化建议、创建测试用例
显著提升了软件开发效率和质量

增强型 搜索引擎

能直接回答复杂问题、进行信息摘要、执行数据分析，绘制图表
朝着提供一站式信息解决方案的方向发展

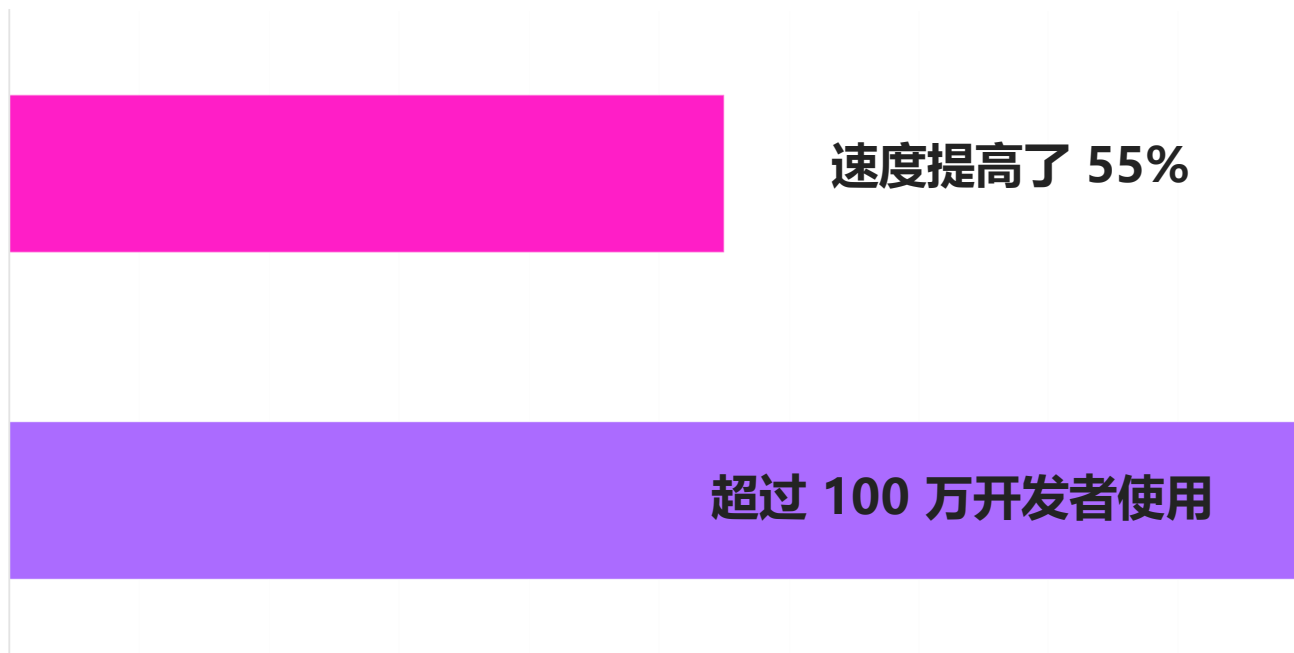
人机交互 革新

交互更加自然、直观，驱动着虚拟助手的能力提升
重塑人与计算机的交互方式



科技行业—案例

GitHub (微软子公司)使用大模型在集成开发环境 (IDE) 中为程序员提供代码建议、自动补全、甚至生成完整函数。





生命与健康

辅助诊断

通过处理海量患者数据（电子病历EHR、医学影像等），辅助医生进行疾病诊断，提高诊断的速度和准确性

药物研发

预测候选化合物的活性、毒性、药代动力学特性
加速药物筛选和研发进程

健康助手

自动监控、分析金融交易，生成合规报告，审查合同
确保满足监管要求，降低合规成本和风险



生命与健康—案例

NVIDIA BioNeMo：加速生成式药物研发的云服务平台
提供了预训练的生物学大模型，加速药物靶点识别、分子生成和蛋白质工程等任务

模型名称	功能描述	特点
ESMFold	从单个氨基酸序列预测蛋白质结构	比AlphaFold 2快得多，在某些情况下精度略低 适合高通量筛选，
MegaMolBART	生成新的小分子结构	助力先导化合物的发现

全球领先的生物制药公司如安进 (Amgen) 正在使用BioNeMo平台来加速其药物发现流程，能够在几周内完成过去需要数月甚至数年才能完成的计算任务



文本创作

大模型辅助作家和编剧进行创意构思、情节设计、对话打磨
甚至直接生成部分内容

图像与视频生成

根据文本描述创作图像、艺术作品、短视频片段或动画。在游戏开发中用于快速生成纹理、概念图等美术资源

游戏引擎

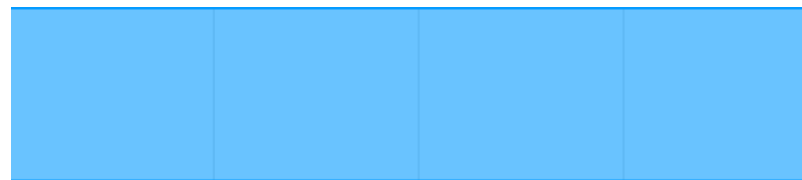
赋予游戏中的非玩家角色（NPC）更自然、动态的对话能力
实时生成游戏环境、事件甚至玩法



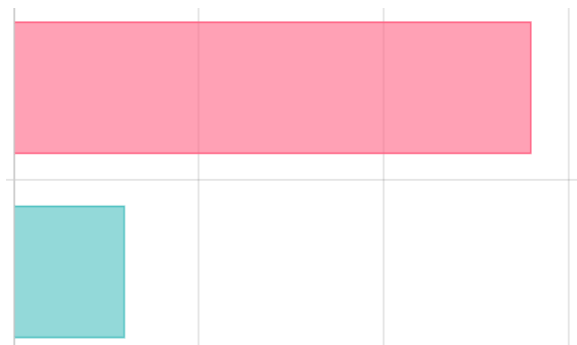
文化娱乐—案例

网易游戏在《逆水寒》手游中广泛运用AI大模型技术，
打造了一系列智能NPC，极大提升了游戏的互动性和沉浸感

与DeepSeek、阿里通义大模型、百度文心大模型、MiniMax-abab大模型、月之暗面-Kimi大模型以及字节跳动豆包大模型等热门国产AI大模型达成合作，共同组成了游戏中的智能NPC女团，由9位NPC少女组成，她们将陪伴玩家在江湖中展开冒险



研发效率
提升30%



任务脚本生成时间
从14天缩短至3天



科学发现

材料科学

预测新材料的性质，辅助设计具有特定性能的材料，解读光谱等实验数据

加速科研流程

在物理、化学、生物等多个基础科学领域，处理大规模数据集，从海量文献中提取和综合关键信息，辅助假设生成和实验设计

AI for Science

从辅助工具转变为半自主的科研主体
能够推动发现并加速所有科学领域的研究进程



科学发现—案例

中国科大机器化学家平台：采用机器智能去查找和阅读文献，从海量研究数据中汲取专家经验，调度2台移动机器人和15个自主开发的智能化学工作站，完成高通量合成、表征、测试的化学实验全流程

智能阅读16000篇论文并自主遴选出5种非贵金属元素，融合2万组理论计算数据和207组全流程机器实验数据，建立了理实交融的智能模型，指导贝叶斯优化程序从55万种可能的金属配比中找出最优的高熵催化剂

传统方法

传统的“炒菜式”遍历搜索方法效率低下，需要检查每一个可能的组合...

1400年

机器化学家

发挥其数据驱动和智能优化的优势.

5周



智能教育

个性化学习

实时分析学生的学习表现和行为数据，提供定制化的学习路径、资源推荐和反馈，满足差异化学习需求

智能辅导系统

充当虚拟教师或辅导员，解答学生疑问，解释概念，提供针对性练习，并给予即时反馈

自动化评估

自动批改学生作业，提供详细评估报告和反馈
减轻教师负担



Duolingo 的 Duolingo Max, 利用大模型提供了两项核心功能: “解释我的答案” (Explain My Answer) 和 “角色扮演” (Roleplay)

解释我的答案

对用户犯的错误进行详细的语法解释, 帮助用户**深入理解错误原因**

不仅指出错误所在, 还会提供**相关语法规则的解释、常见错误类型分析以及正确用法的示例**
让学习者从根本上理解并掌握知识点

角色扮演

提供与**AI角色进行开放式对话**的机会, 模拟真实场景下的语言交流

从餐厅点餐到商务谈判, **多种生活和工作场景**
AI会根据回答实时调整对话难度和方向, 创造**沉浸式的语言学习环境**

Outline

- 自然语言处理
- 大模型的原理与技术
- 大模型的核心技术进展
- 大模型应用
- 未来展望

人工智能的目标

➤ L1: 基础对话者

- **对话与交互能力:** 当前(如GPT-4),能流畅的对话
- **特点:** 以对话为核心, 提供信息、解答问题、辅助创作等

➤ L2: 推理者

- **推理与问题解决能力:** OpenAI-O1接近L2。 PhD-Level。
- **特点:** 像人类一样分析问题、制定方案, 甚至在某些领域超越人类

➤ L3: 智能行动者

- **自主决策能力:** AI能独立思考并根据复杂情境采取行动(接近人类)
- **特点:** 不仅能够思考, 还能在真实世界中执行决策, 实现人机协同

➤ L4: 创新者

- **创造能力:** AI能进行创造性思维和协助人类进行发明和创造, 推动科技进步
- **特点:** 具备创新思维和创造力, 能够提出新的想法和解决方案

➤ L5: 组织者

- **AGI:** AI能执行和组织人类所有工作, 标志着真正的人工通用智能的实现
- **特点:** 具备全面的人类智能和组织能力, 能够管理和优化整个社会的运行。

OpenAI's 5 Step to AGI

Level 1	Chatbots, AI with conversational language
Level 2	Reasoners, human-level problem solving
Level 3	Agents, systems that can take actions
Level 4	Innovators, AI that can aid in invention
Level 5	Organizations, AI that can do the work of an organization

未来发展方向

- **具身智能—物理与数字融合**

- **医疗护理：**康复机器人通过触觉传感器与多模态感知技术，提供个性化康复方案
- **家庭服务：**人形机器人集成情感交互与家务能力，实现烹饪、清洁等场景服务



未来发展方向

科学（研究/发现）范式: LLM4Science



第1范式：经验科学

观察现象
经验

1600年

第2范式：理论科学

理论模型
牛顿定律、电动力学方程等
1950年



第4范式：数据驱动科学

大数据
数据建模、分析、挖掘
2022年



第3范式：计算科学

数值计算
模拟

2010年



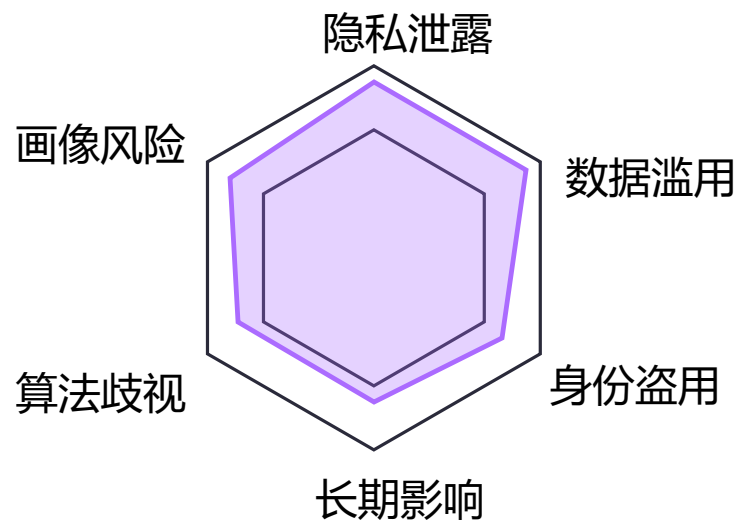
第5范式：智能驱动科学

大模型、智能体
自动化科学研究

2023年

风险与安全

► AI数据泄露



— 数据安全漏洞可能导致：

- 隐私泄露→ 骚扰/诈骗风险
- 数据滥用→ 算法歧视
- 画像偏差→ 发展机会限制

— 未成年人的数据足迹将影响：

- 就业背景调查
- 社会信用评估

风险与安全

风险：生成内容风险



1. 内容可靠性风险

大模型输出存在事实性错误风险，可能造成错误知识的系统性传播。

这种错误在教育教学、社会生活、政务管理中尤为危险



2. 价值取向偏差

算法训练数据中隐含的社会偏见会通过模型输出被放大。

这种偏差对人民价值观塑造 构成潜在威胁



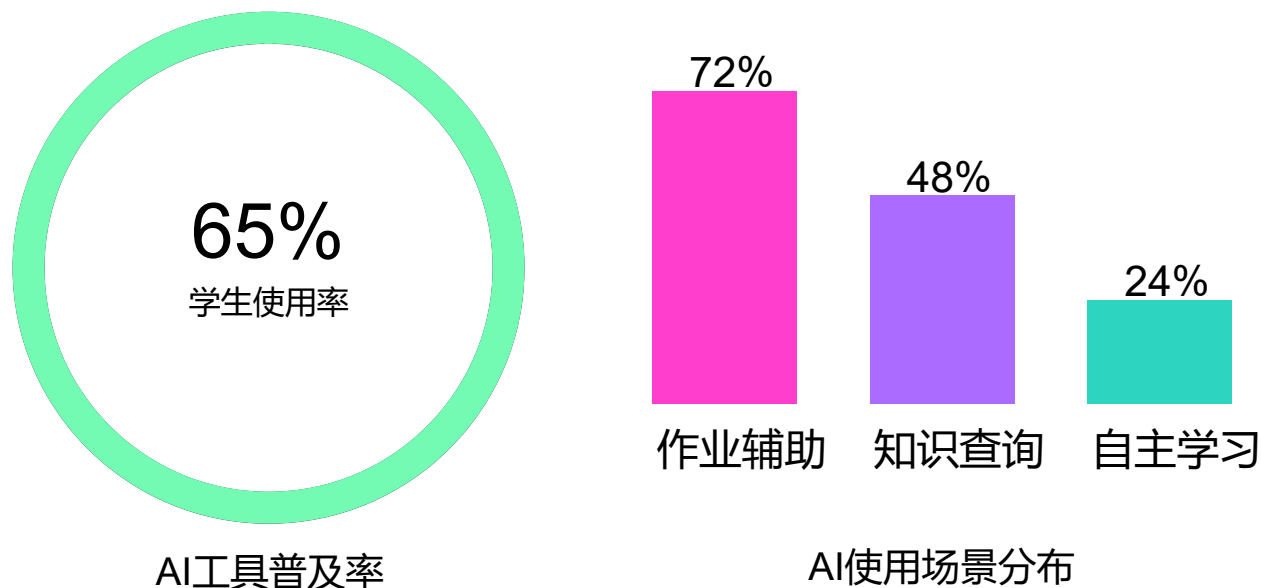
3. 认知框架影响

用户批判性思维尚未完全形成，可能不加批判地接受人工智能输出内容。

对青人民思维发展造成不可逆影响

风险与安全

AI依赖度分析



Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task[△]

Nataliya Kosmyna ¹ MIT Media Lab Cambridge, MA	Eugene Hauptmann MIT Cambridge, MA	Ye Tong Yuan Wellesley College Wellesley, MA	Jessica Situ MIT Cambridge, MA
Xian-Hao Liao Mass. College of Art and Design (MassArt) Boston, MA	Ashly Vivian Beresnitzky MIT Cambridge, MA	Iris Braunstein MIT Cambridge, MA	Pattie Maes MIT Media Lab Cambridge, MA

研究人员指出，相比于“全靠自己动脑”的基准组，使用搜索引擎的那组参与者大脑的 dDTF 连接强度下降了 34% 到 48%；而使用大语言模型（LLM）的那组，下降幅度更大，最多达到了 55%。

DeepSeek变革引发了对学生学习主动性可能被弱化的担忧，如果学生过度依赖人工智能，可能会逐渐失去自主学习的动力和能力，这不仅违背了教育的初衷，还可能对学生的长远发展产生不利影响。

负责任的人工智能

“人在环路” (Human-in-the-Loop)

无论AI多么智能,人类研究者必须始终处于决策和监督的核心位置。AI是强大的工具,但最终的科学判断、结论的采纳和成果的责任,必须由人类承担

常怀敬畏之心 (Maintain Humility)

科学探索的道路充满未知, AI模型本身也存在局限性和不确定性。对技术保持谦逊, 对科学保持敬畏, 是避免傲慢和错误结论的前提

新时代的“智能增强者”

去体验:



不要只听别人说，亲自去试用一个AI工具。

去提问:



练习我们今天学到的方法，
与AI进行一场有深度的对话。

去创造:



用AI辅助你，完成一件你以前觉得
很难或没时间做的小事。

欢迎来到人机协同的新时代。 未来，在自己手中。